Machine learning & semantic web technologies for cancer care

MACHINE LEARNING & SEMANTIC WEB TECHNOLOGIES FOR CANCER CARE

DISSERTATION

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus
Prof. dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public on

THURSDAY 9TH APRIL 2020
AT 14:45

by

Geetha Mahadevaiah

PROMOTOR

Prof. dr. ir. A.L.A.J. Dekker

CO-PROMOTOR

Dr. L. Wee

ASSESSMENT COMMITTEE

1. Prof. dr. J.F.M. Smits (Chair)

2. Dr. H. Hofstraat, Philips Research, Eindhoven

3. Prof. dr. M.S. Hoogeman, Erasmus MedischCentrum Rotterdam

4. Dr. R. Fijten

Hartelijk dank aan de mensen
die geholpen hebben met mijn
promotieonderzoek

Special thanks to all the people
involved in completing my
doctoral research

CONTENTS

7 CHAPTER 1
General Introduction and Outline of the Thesis

17 CHAPTER 2
Automating Data Mining of Medical Reports

39 CHAPTER 3
De-identification of Protected Health Information PHI from Free Text in Medical Records

57 CHAPTER 4
Authorization Framework for Medical Data

81 CHAPTER 5
Semantic Representation of Radiotherapy Data for Effective Data Mining

97 CHAPTER 6
An Approach Toward Automatic Classification of Tumor Histopathology of Non–Small Cell Lung Cancer Based on Radiomic Feature

109 CHAPTER 7
Fractal Analysis in Histology Classification of Non-Small Cell Lung Cancer

123 CHAPTER 8
Cloud Based Big Data Platform for Image Analytics

141 CHAPTER 9
Machine and Deep Learning Based Clinical Decision Support in Modern Medical Physics: Selection, Acceptance, Commissioning and Quality Assurance

163 CHAPTER 10
Discussion and Future Prospective

175 CHAPTER 11
Summary

185 CHAPTER 10
List of Publications

CHAPTER 1

GENERAL INTRODUCTION AND OUTLINE OF THE THESIS

Geetha Mahadevaiah

GENERAL INTRODUCTION

The healthcare industry is in the throes of transformation. Hospitals and clinicians are demanding increased utility from technology to drive down costs, increase efficiency, transparency and deliver better care. There are many societal and environmental changes triggering this transformation, such as an aging population, better informed patients, shortage of trained clinicians and specialists, pervasive social media. Latest technologies such as Semantic Web and Artificial Intelligence can provide solutions to meet the hospitals and clinicians demands.

Gone are the days when the clinician relied on only the stethoscope and physical examination to diagnose health conditions, barring a few routine ailments such as a viral or bacterial infections. Clinicians now rely on the data from machines such as ECG, X-Ray, CT, MR and histopathology investigations to make their diagnosis.

Diagnostic imaging, EMR and clinical decision support systems help clinicians to diagnose, treat and track patients through their journey to recovery. Though these systems are extremely valuable in providing insights, trained technicians, specialists are required to interpret the images and prepare reports.

Studies and interviews with clinicians highlight the distraction introduced by computers and machines in the workflow [1][2]. Valuable time of the physician is spent in entering, reviewing and reporting via a computer, instead of a meaningful and much required dialogue with the patient. Due to the time pressure and complex user interfaces, the captured data is incomplete and prone to errors [3].A statement from Jones Shawn, MD FACS, "I went into medicine to work with people and not to be a data entry clerk", nicely enunciates the issue.

Providing the right information at the right time for the clinicians to make the right decision without overburdening the care providers is a challenge that many researchers, corporate entities and healthcare service providers are attempting to solve. Clinical decision support systems (CDSS) are software applications that analyze healthcare data and might present key insights to the clinicians for better decision-making and solve part of this challenge.

This is most pertinent in the area of cancer diagnosis and treatment. The field of medicine has made significant strides in the area of cancer diagnosis and treatment. The survival rates of cancer patients has improved significantly over the past decades, the reason being early detection and improved treatment strategies. Cancer care is a rapidly evolving, complex, multidisciplinary field, requiring collaboration and information to aid decision making.

During the care cycle in hospitals, vast amounts of data are generated. There are many types of data generated in the hospital setting, such as electronic medical record, image data, waveform data, pathology reports, prescription etc. Broadly, we can categorize these into two major types of data: image data and text data. Image data is obtained from medical equipment such as X-Ray, CT, MR, Ultrasound and ECG. Electronic Medical Records (EMR) and various reporting systems generate text data.

Various technologies have emerged in past few years for sharing large amounts of data, both image and text information, across multiple institutions. Hospitals and healthcare solution providers are leveraging these technologies and integrating them –amongst others - in Picture Archiving and Communication System (PACS) solutions.

The availability of meaningful data is a vital ingredient for an effective clinical decision support system. The latest artificial intelligence technologies such as Deep Learning, depend on large amounts of Findable, Accessible, Interoperable and Re-useable (FAIR) data[11] to create algorithms. These algorithms or models are the soul of the clinical decision support systems.[10]

Medical Text Data Mining

There is a large amount of information available in various medical reports. However, the actual information gleaned from the data trove is miniscule. The reasons for this are manifold, for example non-availability of data in a format, which is conducive for machine assisted mining. Typically, this data is in un-structured format making searching difficult.

Data is locked within hospital departments and not easily available for research or other purposes, also due to privacy and security aspects. Added to this is the difference in nomenclature and medical terms across clinicians, hospitals and countries.

Human intervention is necessary at each step to derive meaning from the data embedded across the various computers and machines in a hospital. Lack of standards makes it difficult to transfer data from one system to another even within the same hospital. Due to issues with interoperability and other factors, sharing of data and information across clinical establishments is a non-trivial task.

The Chapters 2-4, detail the challenges and probable solutions towards extracting and storing context aware information, enabling auto mining while adhering to the security and privacy principles.

In Chapter 2, a detailed description of latest AI based technology to extract meaningful data and store the derived information in Semantic Web technology format, a triples database, for automatic mining of information, is provided.

Security and Privacy of Medical Data

As per the USA HIPPA act and its equivalent in other jurisdictions, Protected Health Information (PHI) from all medical data should be de-identified before the data is shared with researchers or others. There are many techniques and methods to automatically de-identify PHI information and remove or mask it from medical data. The latest techniques are based on 'deep learning" methods, such as NeuroNER, wherein the algorithm is trained on a large corpus of medical text data. The robustness of the algorithm depends on the variety and completeness of the data used for training. A description of the challenges in de-identification, comparison of existing techniques and improvements to the NeuroNER model to identify Indian origin names is provided in Chapter 3.

Extraction of meaningful but anonymized data from medical images and storing it in an RDF repository, enables data mining across multiple institutions. However, this data needs to be secure and the consent of the patient has to be obtained for further processing. Chapter 4 describes a framework developed on RDF data repositories to empower patients to manage the consent and access rights on their data at a granular level.

Mining of Image Data

Vital diagnostic, treatment planning and progress information is stored in DICOM files.

DICOM (Digital Imaging and Communications in Medicine) is the international standard followed by different medical device manufacturers to transmit, store, retrieve, print, process, and display medical imaging information. DICOM facilitates interoperability and communication between medical equipment manufactured by different companies. All the images generated by CT, MR, X-Ray and Ultrasound machines are stored in DICOM format [12].

In addition to the diagnostic images from the modalities stored as DICOM objects, extended DICOM supports standardization of radiotherapy therapy information. A radiotherapy image is stored as DICOM RT Image, the dose distribution is stored as DICOM RT Dose object, the therapy planning information is stored as a DICOM RT PLAN and the target volumes and organs-at-risk information is stored in RT STRUCT[7] [8].

Though DICOM promotes vendor neutrality and improves interoperability tremendously, there are still vendor specific proprietary implementations. Valuable information is stored in proprietary format or deep within a DICOM structure, making it difficult for data mining and insight gathering. DICOM data is not easy to mine, particularly not for extracting longitudinal information on patients for clinical decisions and research purposes. The techniques and methods to publish the data in a DICOM RT Structure in RDF, which can be easily mined for information is explained in Chapter 5.

A promising technology enabling computers to comprehend the data published on the web and supporting trusted interactions is the Semantic Web. Semantic Web technologies such as RDF (Resource Description Framework) are standardized by the World Wide Web Consortium (W3C), for sharing data and meaning beyond boundaries [5]. W3C is a global community of member organizations, full time staff and public participants all developing standards for the web. The leaders of the consortium are the inventor the Web, Tim Berners-Lee, Director and Jeffery Jaffe, CEO.

Quote from Tim Berners-Lee: " I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize."[4]

Data in a format comprehensible by a machine is the essence of the Semantic Web. The RDF provides a simple mechanism to link the subject and predicate by a relation, known as a triple. This description is made unique by means of a Unique Resource Identifier (URI) [6].

In Chapter 5, Semantic Web technologies are used to convert DICOM data into RDF and query these data. Additionally, Natural Language Processing was used to make it easier for non-Semantic Web experts to construct queries.

Radiomics

There is a huge amount of data captured during Radiology imaging. Radiologists study the images in depth and annotate the relevant portions of an image set for diagnostic reporting. The most common features used are the count, location, size and volume of the findings. Radiomics is a fast advancing medical field, which aims to extract more information from the images directly with algorithms [9][13], for clinical decision making.

Diagnosis and decision making might be improved when the data from imaging can be used to derive histopathology information. Chapter 6 describes a Radiomics study to extract the histopathology subtype from image data. This study has shown that the additional Radiomics information improves the accuracy of detection of non-small lung cancer by 20% as compared to the traditional approach of just considering volumetric and shape based features.

Typically, tumors have different shapes and curvatures. The aggressiveness of the tumor spread might be assessed from those characteristics. Fractals are a mathematical method for quantification of irregular patterns. Chapter 7 describes a study in Radiomics, wherein fractal dimensions for small cell lung cancer classification, improved accuracy by 8%.

Richer information, in terms of diversity, quantity and quality, vastly improves decision-making. Thereby, data analytics on information from multiple institutions are expected to result in superior clinical decision support systems (CDSS). A framework, leveraging big data technologies to perform medical image data analytics across multiple institutions, while addressing the challenges of security, privacy and scalability of the solution, is described in Chapter 8.

Clinical Decision Support System

Advances in technologies such as artificial intelligence and machine learning combined with the availability of large volumes of data is leading to an increased adoption of CDSS in hospitals.

Latest technologies such as deep learning are revolutionizing the way CDSS are developed, validated and deployed in hospitals. Machine learning is a technique to get the computers to act without explicit programming. Deep learning is a subset of machine learning. Deep learning requires large amounts of annotated and curated data and right configuration of hardware. This is a continuous learning system, wherein the system improves and updates as it gains new information from the latest data feed.

There is a rapid development of a vast number of decision support tools based on the state of the art artificial intelligence technologies by the industry and academia. Thereby there is an urgent need for hospitals and manufactures of machine learning software to fine-tune their processes of development, implementation and validation for a successful deployment of the clinical decision support systems [10]. A review of the clinical decision support systems based on the latest technologies and the inherent challenges in hospital deployment are laid out in Chapter 9.

REFERENCES

1. *Electronic Health Record Management: Expectations, Issues, and Challenges: Mathai N*, Shiratudin MF and Sohel F; Mathai et al., J Health Med Informatics 2017, 8:3; DOI: 10.4172/2157-7420.1000265*
2. *Resistance to Electronic Medical Records : A barrier to improved quality of care ; David B Meinert ; Southwest Missouri State University, Springfield, Missouri, USA. Issues in Informing Science and Information Technologies.*
3. *Electronic Medical Records : Promises and Problems; William R Hersh; Journal of the American Society for Information Science; 46(10):772-776 1995*
4. *https://en.wikipedia.org/wiki/Semantic_Web*
5. *https://www.w3.org/standards/semanticweb/*
6. *https://www.w3.org/wiki/URI*
7. *DICOM demystified: A review of digital file formats and their use in radiological practice. Graham, R.N.J. et al.Clinical Radiology, Volume 60, Issue 11, 1133 – 1140*
8. *DICOM-RT and Its Utilization in Radiation Therapy Maria Y. Y. Law, Brent Liu; RadioGraphics Vol 29, No. 3; Published Online: May 1 2009. https://doi.org/10.1148/rg.293075172*
9. *Radiomics: Images Are More than Pictures, They Are Data Robert J. Gillies, Paul E. Kinahan, Hedvig Hricak: Nov 182015 https://doi.org/10.1148/radiol.2015151169*
10. *Artificial Intelligence and Machine Learning in Software as a Medical Device; https://www.fda.gov/MedicalDevices/DigitalHealth/SoftwareasaMedicalDevice/ucm634612.htm*
11. *Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific datamanagement and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).*
12. *Bidgood WD Jr, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. J Am Med Inform Assoc. 1997 May-Jun;4(3):199-212. doi: 10.1136/jamia.1997.0040199. PubMed PMID: 9147339; PubMed Central PMCID: PMC61235.*
13. *E.D'Arnese, G.W di Donata, E.del Sozzo, M.D. Santambrogio, "Towards an Automatic Imaging Biopsy of Non-Small Cell Lung Cancer", 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2019.*

CHAPTER 2

AUTOMATING DATA MINING OF MEDICAL REPORTS

Geetha Mahadevaiah
Dinesh M.S.
Amogh Hiremath
Vani Agarwal
Ponnurangam Kumaraguru
Andre Dekker

International Journal of Computer Science and Technology (IJSCT)
VOL. 01, NO.2, MARCH 2019
DOI: 10.5121/IJCST.2019.041012 1

ABSTRACT

Medical reports contain large amounts of clinical information which is not easily mined due to its unstructured and free flowing format. In addition, medical terminology is context sensitive and varies between entities.

In this study, a method to convert data from unstructured medical reports into structured format and techniques to identify cancer cases including classification on the basis of its type, stage, occurrence in an organ, are devised.

Different NLP techniques were evaluated for feature extraction. A machine learning based algorithm for automatic information extraction was developed. The system performs well for malignant/benign cancer classification with 0.89 F1-score.

Semantic structure of reports in the form of ontology was developed, enabling machine comprehensible data storage and retrieval of semantic information. For illustration purpose, the semantic representation of lung cancer cases with mapping from clinical reports is shown.

Keywords

CLINICAL REPORTS
REPORT CLASSIFICATION
CANCER CLASSIFICATION
RDF FILES
ONTOLOGIES
NATURAL LANGUAGE PROCESSING

1.0 INTRODUCTION

Medical reports contain detailed clinical information about a patient's medical condition. They include patient information like findings, impressions, demographics, past medical history of the patient, brief hospital course, diagnosis etc. A large proportion of these reports are in unstructured free-text format. During the medical journey of the patient, based on the need, medical practitioners access information in the medical reports at various points. During this process, medical practitioners or clinicians have to read many reports to gain insights about the patient's condition. There is a real possibility that the doctors might miss critical information which is aggravated by time pressure and resource constraints.

The information in the unstructured text data may provide additional medical insights and therefore are of interest to the clinical research community. An example of information one would like to extract from free-text clinical reports is the disease of the patient, which would allow automatic grouping of clinical reports into various diseases for further study. The discharge summary notes of patients are typically labelled according to ICD (International Classification of Disease) codes [2] .

Since 1967, the World Health Organization (WHO) has developed ICD codes to monitor the incidence and prevalence of diseases, observe reimbursements and resource allocation trends, and to keep track of safety and quality guidelines. In the hospital setting, to perform annotations, technicians have to first classify reports according to their disease types and ICD codes. In the medical field, ontologies such as ICD are commonly used to represent knowledge about symptoms, diseases and treatments. Mapping of the disease information to such predefined ontologies is done manually which requires manpower and is prone to errors.

There is a need for a system that automatically groups medical reports into various diseases to help in the early identification of symptoms and treatment of disease. Semantic web and Natural Language Processing provide methods to convert unstructured data into a structured format and store data with semantic context and in a format that machines can process. Automated processing and identification of information from medical reports might support medical practitioners to derive clinical insights and provide treatment in a faster and more precise manner.

Cancer is among the leading causes of death worldwide[4]. Cancer care is multidisciplinary and an ideal test case for the wider use of electronic health records to manage oncology data and workflows [3]. There is an increase in the occurrence of cancer with related morbidity, thus the need of the hour is to build an automated system for its early detection.

In this work, the authors have devised techniques to process and extract information from medical reports, convert them to a structured format and map disease information in a semantic form and applied it to cancer.

2.0 METHODS

This section provides a brief description of related work in the area of automated medical report annotation. It provides details of the methods and techniques of different approaches and their benefits.

2.1 Related Work

To automate the process of tagging medical reports, previous studies have used many techniques for automatically tagging ICD9 codes to medical reports. Rule based techniques are used to tag reports by using pattern matching and supervised machine learning algorithms [3]. Dublin [4] has used Natural Language Processing techniques to validate pneumonia cases from chest radiographic reports.

Due to unstructured text, there is ambiguity in medical reports and errors such as misspelled words, use of phrases, abbreviations, lexical variations and thus pattern matching methods fail to provide comprehensive results. Rule based systems are similar to the manual annotation of reports. Wang [5] showed that NLP techniques have gained power and competence compared to rule-based techniques.

Machine Learning methods have also been used in Banerjee [6] who mainly focus on word embeddings using an unsupervised hybrid method. Word embeddings

are formed by training Word2Vec [32] on text data. The method proposed by the authors combines word embeddings with a semantic dictionary mapping technique for creating a dense vector representation of unstructured radiology reports. Further, they have applied intelligent Word embeddings to generate embedding of chest CT radiology reports from two health care organizations and utilized the vector representations to semi-automate report categorization based on a clinically relevant categorization related to the diagnosis of pulmonary embolism (PE).

Oberkampf[7] implemented a prototype to demonstrate structured representation of findings from unstructured reports, which allows the successful review of and more efficient comparison among various EHRs. The role of the information model proposed is to define the schema according to which the terminology is used. In previous work, Zillner[8] created a Model for Clinical Information (MCI) that is based on ontologies from the Open Biological and Biomedical Ontologies (OBO) library[9,10]. RadLex [11], and other ontologies are employed as reference terminologies. Further, Oberkampf [12] demonstrated how structured representations of measurement findings can be extracted from free-text radiology reports.

Mabotuwana[13], proposed a measure to determine similarity between two individual concepts extracted from a free text document. They used ontological parent-child (is-a) relationship as matching techniques, as lexicon based comparisons are typically not sufficient to determine an accurate measure of similarity. The addition of semantic context into the document vector space model improves the ability to differentiate between radiology reports of different anatomical and image procedure-based classes. This effect, when leveraged for document classification tasks, can be used for efficient biomedical information retrieval.

Mahadevaiah[14] demonstrated the use of semantic technology constructs to store clinical relevant features from the DICOM (The Digital Imaging and Communications in Medicine) standard , which is widely used in medicine for storing and transmitting medical information especially related to imaging. Natural Language processing were used for mining and retrieval of information. The proposed technique stores the clinically relevant information, from a DICOM RT dataset, as triples in a Resource Description Framework (RDF) repository.

Studies discussed in the above paragraphs describe NLP techniques, to demonstrate automatic text processing. But none of the discussed techniques provide a disease specific classification with the representation of sub-categories of disease using an ontology. The

proposed work here, as shown in Figure 1, builds an automated text classifier which can identify malignant or benign cancer types and its sub classification by processing unstructured text data. This information is represented as an ontology to enable easy retrieval of insights on a patient's disease progression to a clinician.

Medical records from the MIMIC-III (Medical Information Mart for Intensive Care III) dataset[1] were used to design and validate proposed techniques. It contains de-identified medical records of patients suffering from various diseases within the intensive care units at Beth Israel Deaconess Medical Center from 2001 to 2012 in free-text format. The MIMIC-III dataset contains ICD9 codes. The ICD9 labelling of clinical reports was done manually under the expert supervision of doctors, which is a labor-intensive task.

The MIMIC-III dataset contains diverse information. In this work, discharge summaries of patients that are present in free-text format were used (The NOTE EVENTS table in MIMIC-III contains the discharge summary). Rule-based regular expressions techniques were used to convert these reports into structured format.

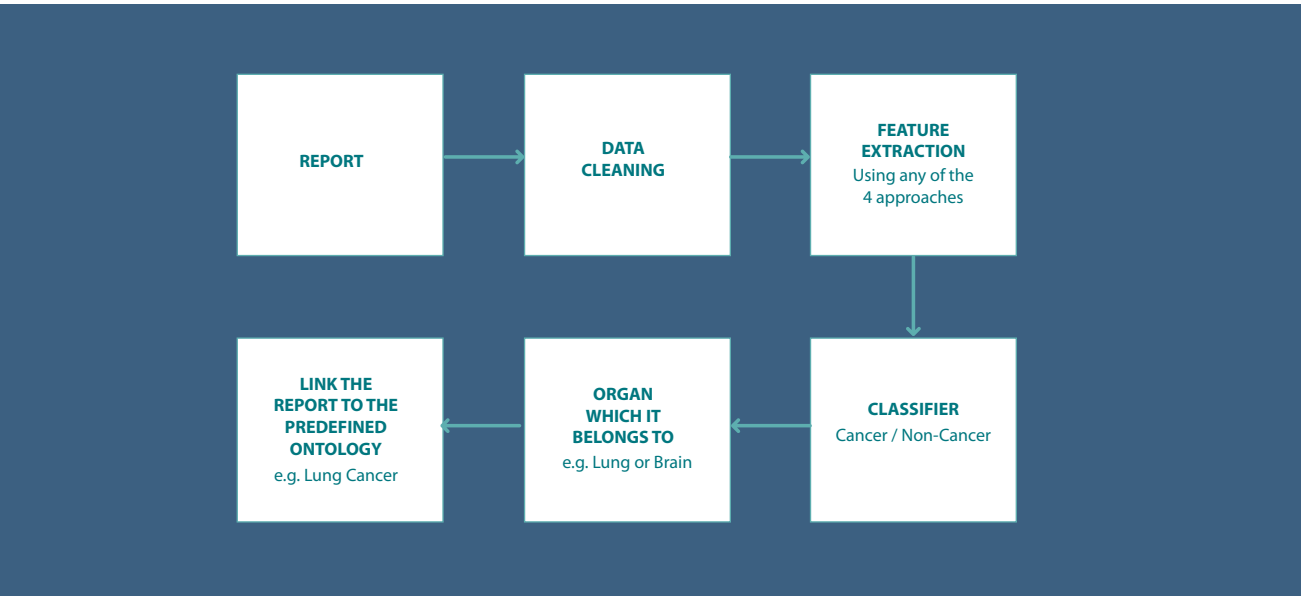


FIGURE 1. ***Overview of the approach used for classification***

2.2 Data and Ground Truth

The sample reports from the MIMIC database, contain discharge diagnoses labels from which the findings can be derived. The corresponding ICD9 codes for the findings are also stored in DIAGNOSES_ICD table of MIMIC-III database and these are used as the ground truth for cancer identification.

In ICD9 code ranges specified by National Cancer Institute, there are 389 ICD9 codes related to cancer [15,16]. The reports were scanned for all the ICD9 codes and 6228 reports with a cancer related ICD code were shortlisted for further processing.

Among the shortlisted 6228 reports, the distribution of sanitized reports based on organ and type of cancer are listed in Table 1 and 2 respectively. Reports were sanitized by removing references to ICD9 codes. Figure 2 shows the overview of the steps followed for ground truth extraction.

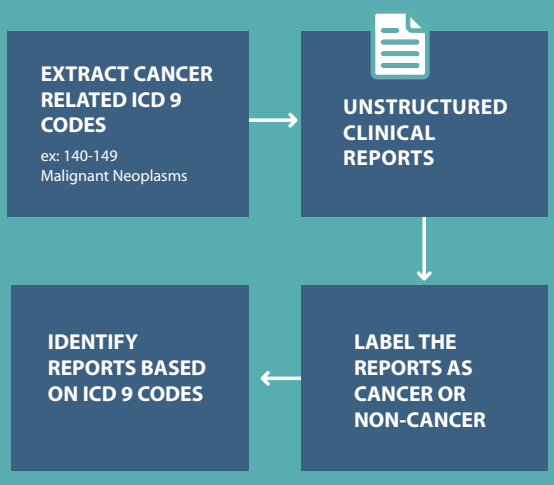


FIGURE 2. ***Ground truth extraction overview***

CANCER TYPE	COUNT
Lymphoma	537
Sarcoma	327
Mastocytoma	70
Carcinoma	57

TABLE 1. ***Cancer Type based count***

ORGAN	COUNT
Head & Neck	1317
Lung	1177
Lymph Node	709
Urinary Tract	376
Colon	305
Prostate	302
Brain	294

TABLE 2. ***Organ based count***

2.3 Data Preprocessing

Unstructured reports were converted into structured format by carefully analyzing discharge summaries and processing various sections like findings, impressions, medical history etc. The processing was done using NLTK - Natural Language Processing Toolkit. This is a Python software library for text processing, such as case conversion, removing special characters, canonizing numerals and tokenizing. The processing steps are described in Figure 3.

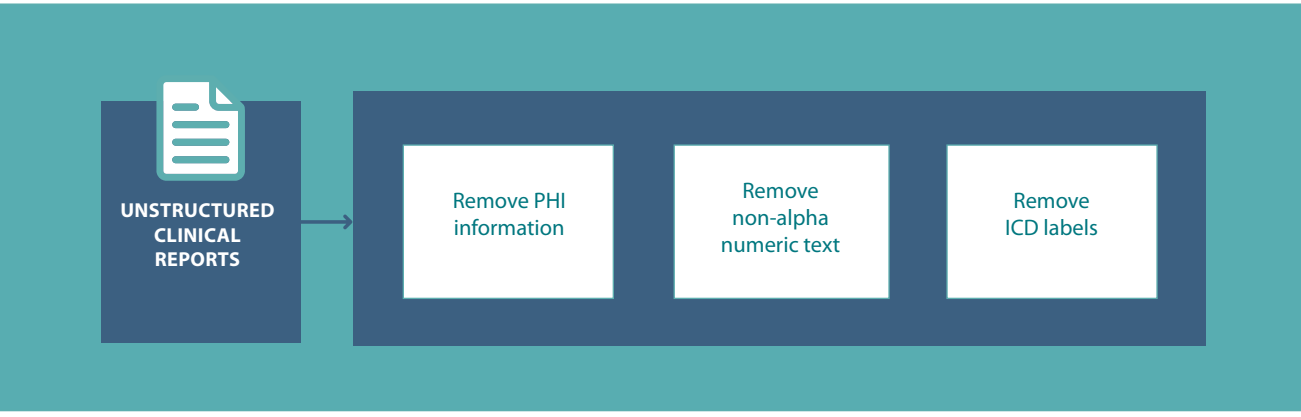


FIGURE 3. ***Data Preprocessing Steps***

Organ classification was based on the ICD9 codes. The discharge summaries were processed and the corresponding ICD9 code of an organ was used to identify the organ mentioned in the report. The dataset thus obtained was sparse as seen in Table 1, hence, up-sampling and down-sampling techniques were used to balance the dataset.

2.4 Feature Extraction

In this work, four approaches for feature extraction namely, 1) term frequency-inverse document frequency (TF-IDF), 2) a Word2Vec model, 3) a combination of TF-IDF and Word2Vec model 4) pre-trained Word Embeddings. Word2Vec is a set of pre-trained models to generate word embedding [21]. TF-IDF is used to weigh the terms based on their frequency of occurrence.

TF-IDF is a measure used to evaluate the significance of a word in a document within a collection or corpus. TF is the frequency of the word in that document and IDF measures the commonness of the word among various documents. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

$$IDF(w) = \log(N / nw)$$

where Nd is the total number of documents and nw is the number of documents that contain word w. To calculate TF-IDF, the first step was to tokenize all the notes, then at later stages the document-word matrices, which stores count of each word (TF) multiplied by the IDF weight, were created.

Word2Vec models [17] are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words to produce word embedding. Word2Vec takes as input a large corpus of text and produces a vector space, typically of several hundred dimensions and similar words are assigned vectors in close proximity. Continuous Bag of Words (CBOW), a feed-forward neural network model that consists of inputs, projection and output layers were used in this feature extraction technique. The inputs are words and the CBOW model predicts the target word based on the context, that is, words that precede and follow the target word. As the corpus to train the Word2Vec model, text notes from MIMIC-III and pre-trained word vectors from PubMed [18,19] were used. PubMed contains more than 27 million records of articles in the biomedical literature and items from the NCBI books database.

2.5 Approach 1 - TF-IDF

To generate a feature vector for a report, TfIdf Vectorizer [20] was trained on clinical terms present in the report, representing clinical terms as n-grams (unigrams, bigrams, trigrams). To extract clinical terms, the Named Entity Recognizer was used. The clinical terms extracted are combination of words or single words like x-ray, lung cancer etc.

The fragmentation of these words into unigrams would result in a loss of meaning of the term lung cancer, therefore bigrams were chosen. Similarly, trigrams were selected for medical terms containing three terms . A TF-IDF fit on clinical terms was performed to obtain the feature vector corresponding to the training reports and this was transformed to the test reports using a TF-IDF fit model. Figure 4 shows the data set used to extract TF-IDF feature extraction and its output format.

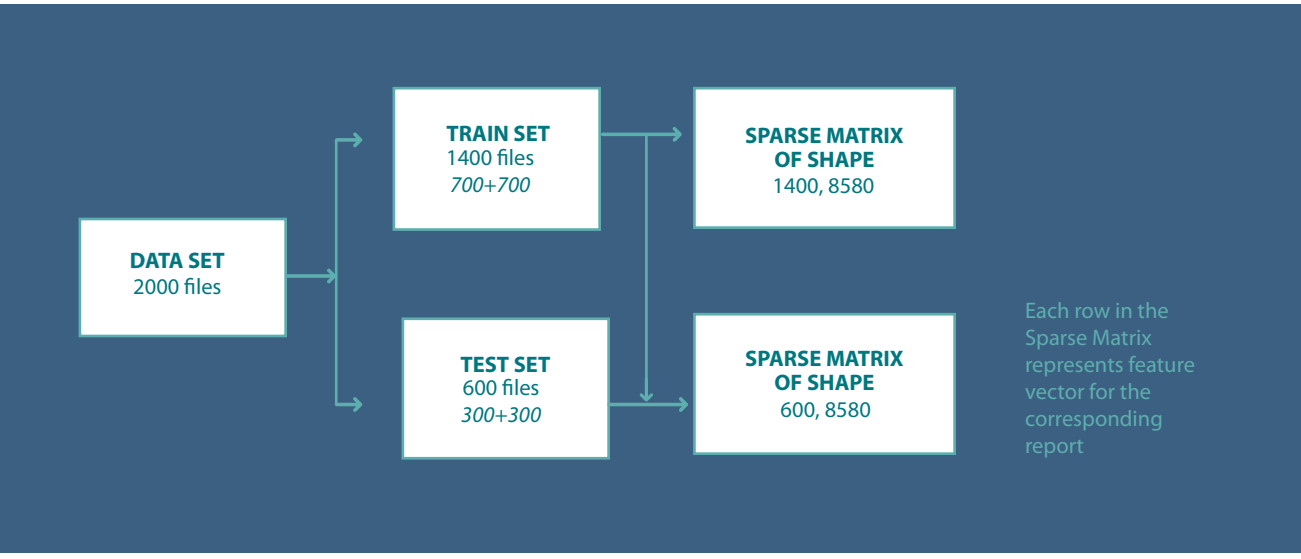


FIGURE 4. ***TF-IDF Medical Terms Extracted Using NER***

2.6 Approach 2 - Word2Vec

Feature vectors corresponding to a report were formed using the following steps:

1. For computing word vectors, text from all the reports were combined, Word2Vec was trained to generate word vectors for all the words in corpus.

2. To obtain vector representations for each sentence in a report, mean average pooling was performed.

3. K-means clustering algorithm was trained to cluster all the sentences.

4. For each report:

i) Sentence vectors were computed.

ii) K-means clustering of these sentences.

iii) Sentences counts in each cluster were used to create a histogram and generate a Probability Distribution Function (PDF).

Figure 5 provides an overview of the method used and a sample word vector is given in Figure 6.

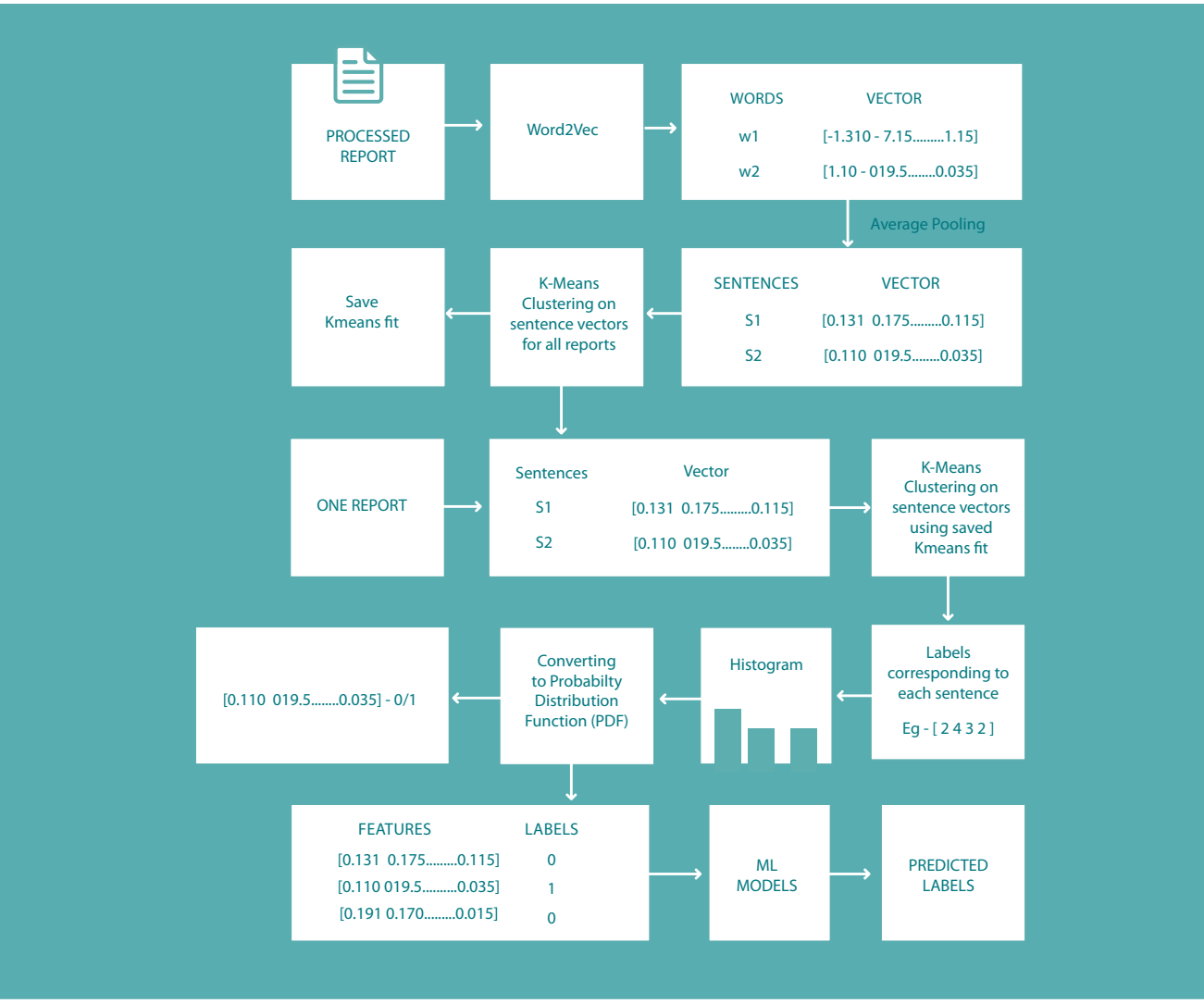


FIGURE 5. ***Word2Vec & K-means feature extraction flow***

```
  Bone:100 features using Word2Vec

array([-0.45568326,   -0.42008284,    0.5879811 ,   -0.5954721 ,    0.242069  ,
        0.04314294,   -0.11443035,   -0.42404965,   -0.48740718,    0.7659249 ,
       -0.5828809 ,    0.7553276 ,   -1.1470004 ,   -0.1757206 ,   -0.18863104,
        0.21828651,   -0.5719981 ,    0.16566484,    0.20187922,   -0.41965342,
        0.37324256,   -0.10080674,    0.01068049,   -0.2813971 ,   -0.2829634 ,
        0.42982036,   -0.18310162,    0.13920984,    0.5431863 ,   -0.65572494,
        0.20396811,    0.00256491,    0.03591614,   -0.5485845 ,   -0.5153154 ,
        0.46385816,    0.22171667,   -0.11218242,   -0.1582615 ,   -0.08939845,
        0.06131884,   -0.3892101 ,    0.23531616,    0.27197105,   -0.5130427 ,
       -0.24943025,    0.10215823,   -0.89288765,   -0.42185077,    0.14471734,
        1.0374857 ,   -0.008071  ,    0.02999489,   -0.29102188,   -0.50396645,
        0.08793559,    0.3439966 ,   -0.34348166,    0.26452625,    0.38907006,
        0.4787299 ,    0.3284534 ,   -0.04414903,   -0.04706165,    0.03913996,
        0.04705186,    0.09013759,   -0.08766278,    0.09889007,    0.27696326,
        0.24847467,    0.21624993,   -0.4111728 ,   -0.08203211,   -0.21875288,
        0.01415944,   -0.00842295,    0.02118845,    0.04208755,   - 0.23537743,
        0.23402141,   -0.01921754,    0.30300575,   -0.0511810 ,   - 0.03348159,
       -0.24223523,   -0.12935163,    0.47220245,   -0.0772469 ,    0.23612452,
       -0.3045119 ,   -0.03456276,    0.18997438,   -0.37388992,   -0.18503377,
        0.32238147,    0.4318316 ,   -0.2188602 ,   -0.46486193,    0.11918008]
     dtype=float 32)
```

FIGURE 6. ***Word Vector – Example (Len = 100)***

2.7 Feature Analysis

Based on a Fisher score analysis of cancer and non-cancer discrimination, final sets of features were selected for classification.

1. Average feature set reduced from [1000x300] to [1x300] is shown in Figure 7a.

2. Using fisher discrimination analysis, out of 300 features, the 10 features which give good discrimination between cancer and non-cancer were selected as the final set of features for classification. Figure 7b shows the discrimination details of the 10 selected features.

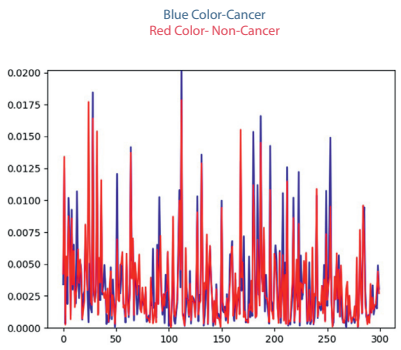


FIGURE 7A. ***Feature Analysis Plot***

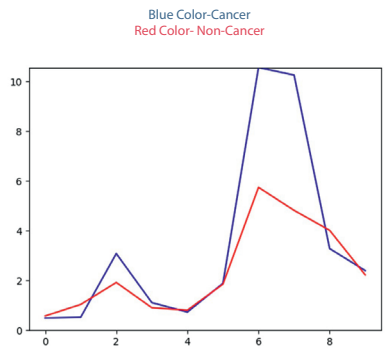


FIGURE 7B. ***10 selected features***

2.8 APPROACH 3
Hybrid Model consisting of both TF-IDF score and Word Vectors

In this work, Word2Vec and TF-IDF features were combined to get stronger features. The following process explains the hybrid model approach.

The processing steps for the hybrid model are:

1. Train TF-IDF on clinical terms to obtain the TF-IDF score. The TF-IDF score was obtained from the Named Entity Recognizer.

2. Train Word2Vec on sentences to obtain word vectors for each word. (As described in Approach 2).

3. Multiply the TF-IDF weight of clinical terms to their corresponding word vector and perform average pooling of these new word vectors to make a sentence vector.

4. Follow K-Means clustering process as in Approach 2.

2.9 APPROACH 4
Using Pre-trained Word Embedding (PubMed & PMC)

In this approach, pre-trained PubMed and PMC word embeddings [18,19] of the Word2Vec model were used and the remaining steps are similar to the Word2Vec and K-means Approach. By this method, features with corresponding labels were obtained.

Different classification models were trained using a grid search approach as Word2Vec model has various hyper-parameters [23] such as, num_features, num_workers, context_size, min_word_count etc. to search over. Among these, num_features and context_size play a crucial role as num_features is the size of a word vector and context_size is the size of context window (8). Experiments were performed with different values of num_features (100, 200, 300, 400, 500, 600, 800) and context_size (5, 10). Also for k-means clustering, different k values were tested, ranging from 100 to 1000.

To build a model for classification into a malignant or benign cancer, we trained different machine learning models, analyzed performance of each model with different hyper parameters and selected the model with best performance. Models used for the grid test

were based on Logistic Regression, Support Vector Machines, Random Forest classifier and Gradient Boosting Classifiers.

Different oversampling techniques [24] have been used to classify cancer occurrence in a particular location/organ. In the our approach, experiments were conducted with Random Over Sampling, Synthetic Minority Oversampling Techniques (SMOTE), and ADASYN at the Word2Vec processing step to increase the sample size.

2.10 Semantic Information Extraction

Semantic Web was used to find self-describing interrelations of data in a form that machines can process. The structured format facilitates storage and information retrieval based on the meaning and logical relationships. Instead of retrieving matching text for a query, the technology permits us to find related text.

The following steps describe the method to build an ontology for lung cancer reports:

1. Extracted semantic information from MIMIC reports and represented them as an ontology structure. Corresponding to each term selected, the relevant predefined ontologies were extracted using the Bio Ontology API [25].

2. Existing ontologies (SNOMED[26], RADLEX [11], LOINC [27]) were used to link medical findings to the ontology structure.

3. Resource Description Framework (RDF) [28] was used to create triples (subject, predicate, object) where the subject is the URI corresponding to ontology, the predicate describes the relationship between subject and the object is either a URI or a string or literal.

4. For building the ontology [29,30], the information in lung cancer reports were converted into the graph structure as shown in Figure 8 (image source [31]). A Python scripts was developed to convert reports to this RDF structure. Various namespaces were added for existing ontologies. For each sentence from a report, triples are formed as shown in example below. These sentences are linked as an RDF graph as shown in Figure 9.

The following are the triples for the example sentence: "Patient1 with small bilateral pleural effusions".

(Patient1, has, findings1)

(findings1, consist, pleural effusion)

(pleural effusion, is, bilateral)

(pleural effusion, effect/size, small)

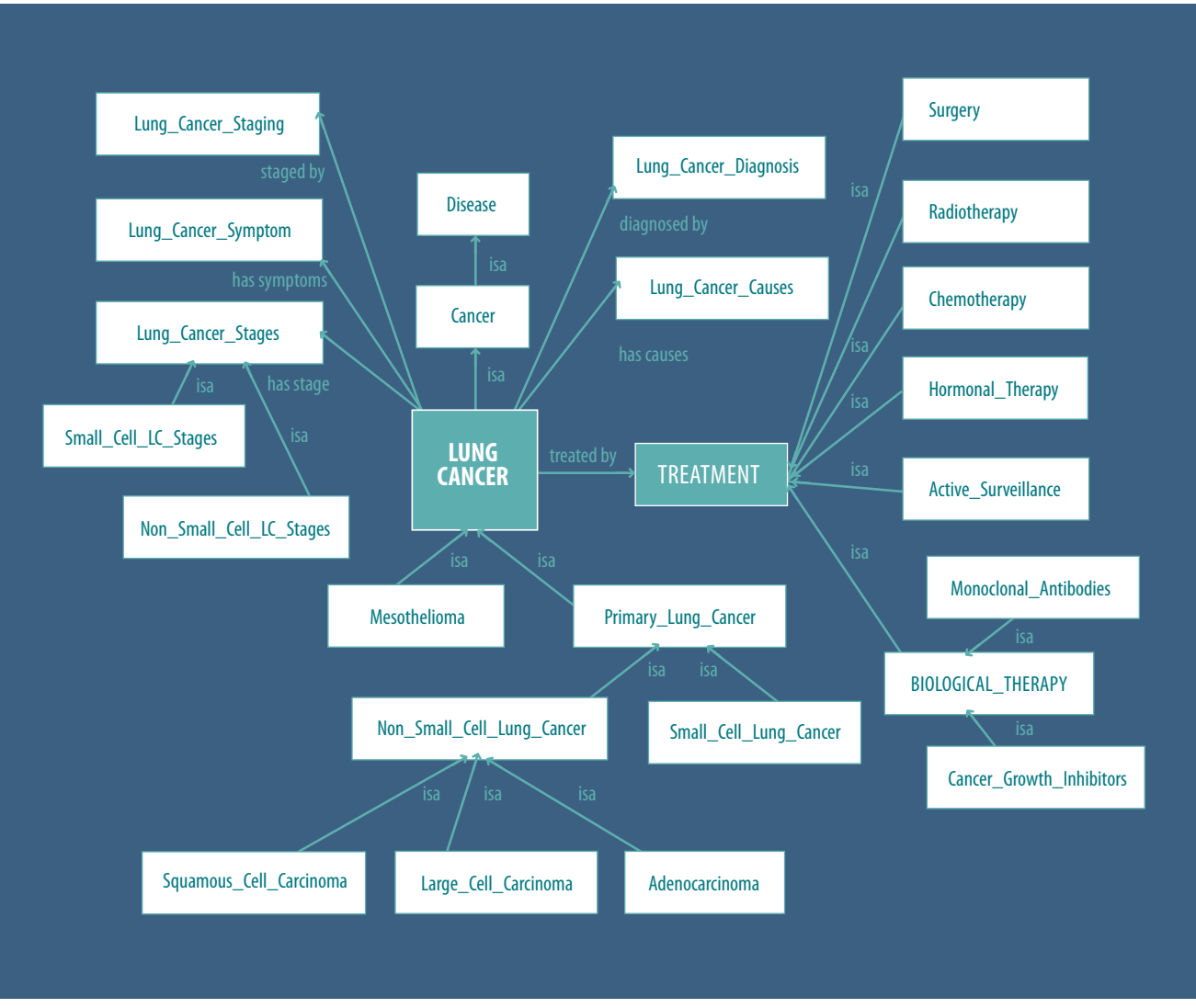


FIGURE 8. ***Lung cancer graph structure***

Ethical considerations

This research did not involve human subjects. Anonymized patient data from publicly available database, such as MIMIC and ICD codes were used to develop and validate the proposed methods.

3. Results

For classification of medical reports into malignant or benign cancer reports, the medical report dataset was prepared for class balance on 12456 samples and the classification results obtained are as shown in Table 3. Further, organ information was extracted from medical reports to map the reports to organ and organ cancer ontologies. For organ based classification of cancer, the dataset created was imbalanced. To overcome this class imbalance problem, oversampling techniques were employed to balance the data and the results are tabulated in Table 4. By analyzing values of different hyper-parameters, the best parameter value for num features are from 300 to 600. A gradient boosting classifier performed well over other classifiers for both TF-IDF and Word2Vec models.

A comparative analysis of different approaches discussed in section 2.5 to 2.9 shows Word2Vec produces better result than a combination of Word2Vec and TF-IDF. TF-IDF and (Word2Vec + TF-IDF) produce almost similar results for gradient boosting classifiers. The results using word embeddings trained on reports improved compared to pre-trained word embeddings such as PubMed and PMC texts. Wikipedia PubMed + PMC articles produces comparable accuracy to Word2Vec but Word2Vec produces an improved F-score as the embeddings are trained on the MIMIC-III reports. This is the general trend observed among all the classifiers on which the experiments were done.

Figure 9 shows the ontology based semantic graph built for the lung cancer reports. This figure contains the lung cancer (adenocarcinoma) graph structure identified in one of the reports on the left and its corresponding RDF structure on the right. The nodes in the RDF structure are id values of a predefined ontology taken into consideration while building the graph. The ids are obtained from the NCBO API [25].

Approach	Classifiers							
	Logistic Regression		SVM		Random Forest		Gradient Boosting	
	F-Score	Accuracy	F-Score	Accuracy	F- Score	Accuracy	F Score	Accuracy
TF-IDF	0.70	0.839	0.85	0.876	0.80	0.817	0.87	0.889
Word2Vec	0.75	0.791	0.87	0.887	0.87	0.852	0.89	0.899
Word2Vec+ TF-IDF	0.69	0.75	0.84	0.87	0.79	0.86	0.84	0.887
PMC pre-trained	0.80	0.823	0.81	0.875	0.87	0.867	0.83	0.877
PubMed pre-trained	0.75	0.83	0.80	0.866	0.75	0.87	0.81	0.878
Wikipedia PubMed + PMC articles	0.78	0.811	0.83	0.87	0.81	0.872	0.80	0.891

TABLE 3. ***Classifier performance for malignant/benign cancer***

Approach	Classifier			
	Logistic Regression		Multilayer Perceptron (4 layers)	
	F- Score	Accuracy	F- Score	Accuracy
Original Dataset with no sampling (IMBALANCED)	0.12	0.27	0.32	0.37
Dataset built by Random (OVER SAMPLING)	0.37	0.39	0.75	0.76
Dataset built using (SMOTE)	0.39	0.42	0.74	0.765
Dataset built using (ADASYN)	0.37	0.416	0.73	0.72

TABLE 4. ***Classifier performance for cancer specific to organ***

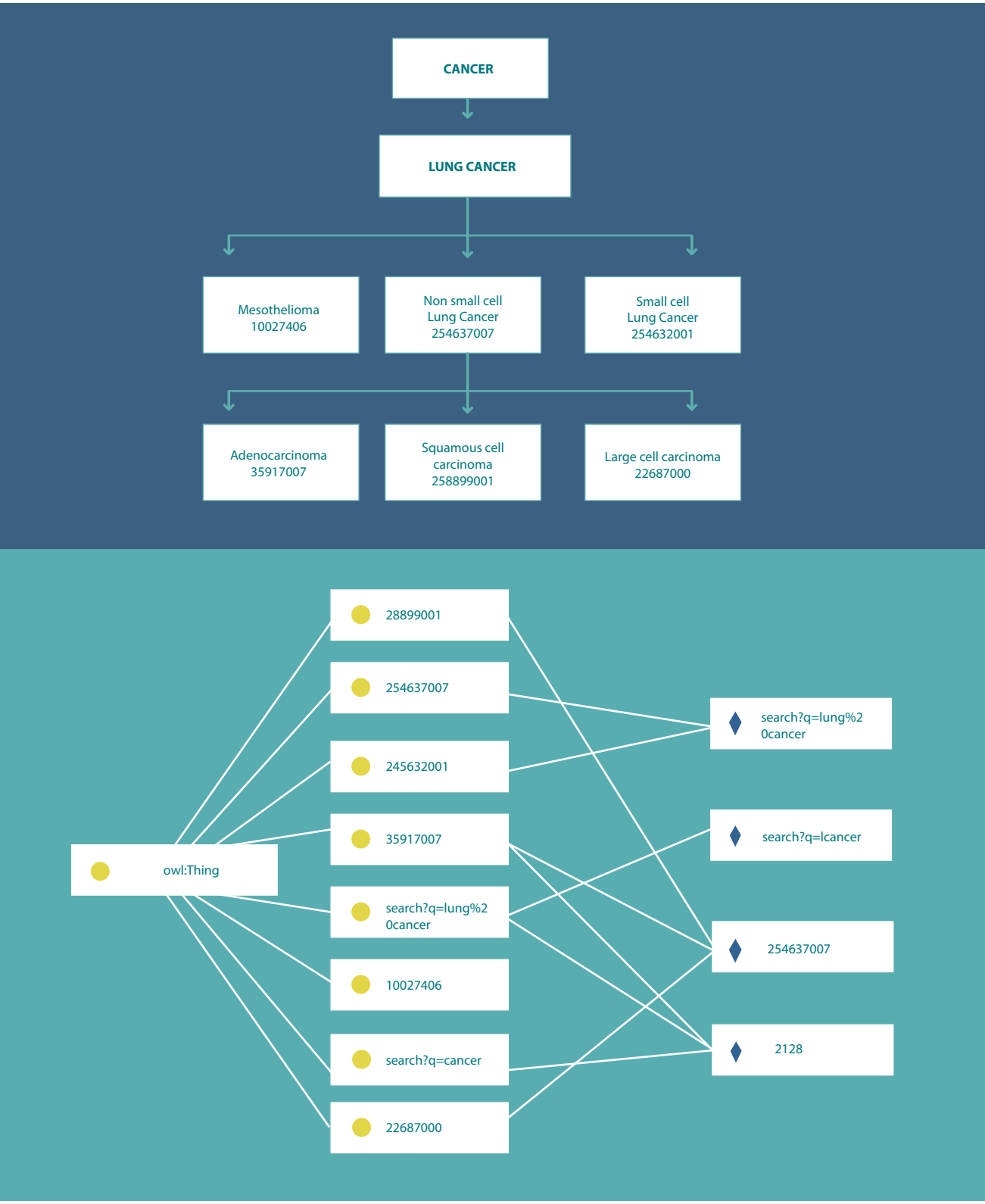


FIGURE 9. ***Semantic representation of lung cancer medical report***

4. Conclusion

In this research work, we experimented with different NLP techniques for feature extraction and developed various machine learning models for classification. An end to end solution from classifying the report to a disease type and mapping disease information to disease ontologies is proposed.

The system performs well for malignant/benign cancer classifications with a 0.89 F-score. Random oversampling method produced a 0.75 F score. Figure 9 shows the semantic representation of the reports for lung cancer and adenocarcinoma ontology. We expect that the proposed techniques will help doctors to extract relevant information from free text reports and brings in efficiency in the clinical workflow.

In the current scenario, a typical medical practitioner has to manually study the historical reports to arrive at a conclusion and there are chances that medical practitioners might miss critical information, aggravated by time pressure and resource constraints. The proposed approach reduces the burden of medical practitioners by improving their throughput by presenting the information in a semantic graph. Visualization of graph data improves readability of information extracted and makes the analysis easy.

The novelty of the proposed work lies in customizing and extending existing NLP techniques to approaches similar to the one used for extracting features from images (histogram). A new methodology is proposed to represent reports into an ontology. This method captures the relationships and links similar concepts. In this work, experiments were conducted by combining TF-IDF and Word2Vec by weighted averaging of Word2Vec vector by TF-IDF weights. The proposed work has significant potential for new clinical applications to targeted cancer treatments.

As a future work, one can further improve the performance of the system by experimenting with deep learning models like RNN and LSTM. Also, the current ontology structure is defined for lung cancer, in a similar way the ontology structure can be extended to different cancer types and further to different disease types as well.

REFERENCES

1. *Johnson AEW, Pollard TJ, Shen L, et al., MIMIC-III, a freely accessible critical care database. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. https://mimic.physionet.org/*
2. *"The International Classification of Diseases, 9th Revision, Clinical Modification" (ICD-9-CM), Sixth Edition, issued for use beginning October 1, 2008 for federal fiscal year 2009. http://icd9.chrisendres.com/*
3. *Alan R. Aronson, Olivier Bodenreider, Dina Demner-Fushman, Kin Wah Fung, Vivian K. Lee1, James G. Mork, Aurelie Neveol, Lee Peters, Willie J. Rogers From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches. BioNLP 2007: Biological, translational, and clinical language processing, pages 105–112*
4. *S. Dublin, E. Baldwin, R.L. Walker, L.M. Christensen, P.J. Haug, M.L. Jackson, J.C. Nelson, J. Ferraro, D. Carrell, W.W. Chapman, Natural language processing to identify pneumonia from radiology reports, Pharmacoepidemiol. Drug Safety 22 (8) (2013) 834–841.*
5. *T Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng 1 , Saeed Mehrabi 2 , Sunghwan Sohn, Hongfang Liu. Clinical information extraction applications: A literature review.*
6. *Imon Banerjee, Matthew C. Chen, Matthew P. Lungren, Daniel L. Rubin, Department of Biomedical Data Science, Stanford University, Stanford, CA, United States Department of Radiology, Stanford University, Stanford, CA, United States. Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort.*
7. *Heiner Oberkampf, Sonja Zillner, James A. Overton , Bernhard Bauer, Alexander Cavallaro , Michael Uder and Matthias Hammon. Semantic representation of reported measurements in radiology.*
8. *Heiner Oberkampf , Sonja Zillner, Bernhard Bauer and Matthias Hammon. An OGMS-based Model for Clinical Information (MCI). In: Proceedings of International Conference on Biomedical Ontology. 2013. p. 97–100.*
9. *Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–5.*
10. *The Open Biological and Biomedical Ontologies Foundry - http://www.obofoundry.org Accessed 31 July 2014.*
11. *RadLex ontology entity, Version 3.11. http://radlex.org Accessed 31 July 2014.*

12. *Oberkampf H, Bretschneider C, Zillner S, Bauer B, Hammon M. Knowledge- based Extraction of Measurement-Entity Relations from German Radiology Reports. IEEE International Conference on Healthcare Informatics; 2014. p.149–154.*

13. *Thusitha Mabotuwana, Michael C. Lee, Eric V. Cohen-Solal Philips Research North America, 345 Scarborough Road, Briarcliff Manor, NY 10510, USA. An ontology-based similarity measure for biomedical data- Application to radiology reports.*

14. *Mahadevaiah G, Soest JV, Dekker A, Udupa N, Rao SV, Kumar YK, et al. Semantic Representation of Radiotherapy data for effective data mining. Proc. Fifth Int. Conf. Adv. Appl. Sci. Environ. Eng. - ASEE 2016 [Internet]. Kuala Lumpur, Malaysia: Institute of Research Engineers and Doctors, USA; [cited 2017 Jun 8]. p. 12–5. Available from: http://www.seekdl.org/nm.php?id=7421*

15. *International Classification of Diseases, 9th Revision, Clinical Modification, Sixth Edition,2014. https://seer.cancer.gov/tools/casefinding/case2014.html*

16. *Mark G. Weiner, M.D., Alice Livshits, Carol Carozzoni, Pharm.D., Erin McMenamin, Gene Gibson, Pharm.D., Alison W. Loren, M.D., Sean Hennessy, Pharm.D., M.S.C.E. Division of General Internal Medicine, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104 Department of Pharmacy Services, Hospital of the University of Pennsylvania. Derivation of Malignancy Status from ICD-9 Codes.*

17. *Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space*

18. *Billy Chiu, Gamal Crichton, Anna Korhonen, "How to Train GoodWord Embeddings for Biomedical NLP" Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 166–174, Berlin, Germany, August 12, 2016.*

19. *Kevin Patel, Divya Patel, Mansi Golakiya, Pushpak Bhattacharyya, Nilesh Birari 1 Indian Institute of Technology Bombay, India ,Dharmsinh Desai University, India, 3ezDI Inc, India. Adapting Pre-trained Word Embeddings For Use In Medical Coding.*

20. *Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al., "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research 12 (2011) 2825-2830 12(Oct):2825−2830, 2011.*

21. *Tomas Mikolov, Karl Chen, Greg Corrado, et al. "Efficient Estimation of Word Representations in Vector Space", arXiv:1301.3781v3 [cs.CL] 7 Sep 2013*

22. *Joseph Lilleberg ,Computer Science Department Southwest Minnesota State University Marshall, MN 56258 USA joseph.lilleberg@smsu.edu Yun Zhu, Yanqing Zhang Computer Science Department Georgia State University Atlanta, Georgia 30302-5060 USA yzhu7@student.gsu.edu, yzhang@gsu.edu. Support Vector Machines and Word2Vec for Text Classification with Semantic Feature*

23. *Gamal Crichton Anna Korhonen Sampo Pyysalo Language Technology Lab DTAL, University of Cambridge {hwc25|gkoc2|alk23}@cam.ac.uk, sampo@pyysalo.net. How to Train Good Word Embeddings for Biomedical NLP*
24. *Imbalanced-learn documentation https://github.com/scikit-learn-contrib/imbalanced-learn*
25. *Bio ontology term search API, http://data.bioontology.org/documentation#nav_search*
26. *U.S. National Library of Medicine, https://www.nlm.nih.gov/healthit/snomedct/*
27. *Unified Medical Language System (UMLS) https://www.nlm.nih.gov/research/umls/*
28. *Decker S, Mitra P, Melnik S. IEEE Internet Computing 2000;4:68–73. Framework for the semantic Web: an RDF tutorial.*
29. *International Classification of Diseases, Version 9 - Clinical Modificationhttps://bioportal.bioontology.org/ontologies/ICD9CM?p=classes\&conceptid=http\%3A\%2F\%2Fpurl.bioontology.org\%2Fontology\%2FICD9CM\%2F782.3*
30. *Lung Cancer Ontology, https://bioportal.bioontology.org/ontologies*
31. *ABDEL-BADEEH M. SALEM, MARCO ALFONSE Computer Science Department Ain Shams University Faculty of Computers and Information Systems CAIRO, EGYPT. Ontology versus Semantic Networks for Medical Knowledge Representation.*
32. *Yoav Goldberg and Omer Levy, "Word2Vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method" arXiv:1402.3722, 2014*

CHAPTER 3

DE-IDENTIFICATION OF PROTECTED HEALTH INFORMATION PHI FROM FREE TEXT IN MEDICAL RECORDS

Geetha Mahadevaiah
Dinesh M.S
Rithesh Sreenivasan
Sana Moin
Andre Dekker [2]

International Journal of Security, Privacy and Trust Management (IJSPTM)
VOL 8, NO 1/2, MAY 2019

ABSTRACT

Medical health records often contain clinical investigation results and critical information regarding a patient's health conditions. In these medical records, along with patient health information, Protected Health Information (PHI) such as names, locations and date information can co-exist. As per the Health Insurance Portability and Accountability Act (HIPAA), before sharing the medical records with researchers and others, all types of PHI information needs to be de-identified or removed. Manual de-identification through human annotators is laborious and error prone, hence, a reliable automated de-identification system is needed.

In this work, various state of the art techniques for de-identification of patient notes in electronic health records were analyzed for their performance. Based on the performance reported in the literature, NeuroNER was selected to de-identify Indian Radiology reports. NeuroNER is a named-entity recognition text de-identification tool developed by Massachusetts Institute of Technology (MIT). This tool is based on Artificial Neural Networks, written in Python, uses Tensorflow machine-learning framework and comes with five pre-trained models.

To test the NeuroNER models on Indian context data such as the name of a person and place, 3300 medical records were simulated. Medical records were simulated by extracting clinical findings and remarks from MIMIC-III data set. For the collection of all relevant Indian data, various websites were scraped to include Indian names, Indian locations (all towns and cities), and Indian Hospital and unit names. During testing of the NeuroNER system, we observed that some of the Indian data such as name, location, etc. were not de-identified satisfactorily. To improve the performance of NeuroNER on Indian context data, a new pre-trained model was added to handle Indian medical reports along with the existing NeuroNER pre-trained model. A medical dictionary lookup was used to reduce the number of misclassifications. Results from all four pre-trained models and the new model trained on Indian simulated data were concatenated and a final PHI token list was generated to anonymize the medical records and obtain de-identified records. Using this approach, we improved the applicability of the NeuroNER system to Indian data and improved its efficiency and reliability. 2000 simulated reports were used as a training set for transfer learning, 1000 reports were used as a test set and 300 reports were used as a validation (unseen) set.

Keywords

DE-IDENTIFICATION
FREE TEXT
PROTECTED HEALTH INFORMATION
MEDICAL RECORDS
RADIOLOGY REPORTS
INDIAN CONTEXT DATA

1.0 INTRODUCTION

Clinical documents contain valuable information regarding disease, medical procedure applied and medication. This has drawn the attention of researchers to explore and extract relevant information from the free form clinical text, such as doctor or nurse notes. However, to use those texts, they have to be de-identified in a way that they give out no personally identifiable information on the patient. During the process of PHI de-identification, it is essential to retain the medical contents of the records to help further research and conserve the value of the record [16].

In the US, the Health Insurance Portability and Accountability Act (HIPAA), guidelines for protecting the confidentiality of health care information was established in April 2003 [7]. Medical records are said to be de-identified, when there is very less likelihood ("risk") of the information alone or in combination with other practically obtainable information, used to re-identify the individual connected with the records.

The 18 specific categories of protected health information (PHI) defined by HIPAA is provided in Table 1[8].The de-identification process should comply with the HIPAA guidelines by obfuscating these specific categories per medical record.

PHI TYPE	NOTES
Names	Both full and partial, but not initials
Locations	All geographic subdivisions smaller than a state, including
Dates	All elements of dates (except years) for dates directly related
Age	All elements of dates (including year) indicative of an age
Telephone numbers	None
Fax numbers	None
Electronic mail addresses	None
Social security numbers	None
Medical record numbers	None
Health plan beneficiary numbers	None
Account numbers	None
Certificate/license numbers	None
Vehicle identifiers	Includes vehicle serial numbers and license plate numbers
Device identifiers and serial numbers	Not restricted to medical devices
Web universal resource locators (URLs)	None
Internet protocol (IP) address numbers	None
Biometric identifiers	Includes finger and voice prints
Any other unique identifying number, code, or characteristic E.g., full face photographic images of full faces, scars or tattoos (and any comparable images).	None

TABLE 1. ***PHI Information***

Given the size of electronic health record databases, the limited number of researchers with access to identified notes [5], and the frequent mistakes of human annotators, manual de-identification is not quite feasible and is expensive in terms of time, effort and cost[4]. Consequently, a reliable automated de-identification system is vital. Failure to accurately "de-identify" a patient's note would put at risk the patient's privacy. Thus the quality and performance of a de-identification system is very important. In this work, we explored various techniques to identify PHI information and then de-identify the same for the purpose of use by researchers [5].

Various state of art techniques for de-identification were analyzed for their performance and NeuroNER was selected for further enhancements of a de-identification system to address Indian context data [20].

NeuroNER is a named-entity recognition tool based on Artificial Neural Networks written in Python and uses the Tensorflow machine-learning framework. It uses bi-directional LSTM (Long Short Term Memory), along with a CRF layer (Conditional Random Field). It has five pre-trained models, which are Conll , I2b2 GloVe SpaCy, I2b2 GloVe Stanford, Mimic GloVe SpaCy and Mimic GloVe Stanford. Among these, Conll was trained on Reuters data which is based on American, European and Asian stock market indices and the other four were trained on medical data. These datasets were prepared from various major sources available using SpaCy and Stanford NER taggers. It also uses GloVe pre-trained token embedding.

Since NeuroNER is based on machine learning, its output and efficiency depend on the kind of data used for model training. In addition, when the tool was further tested, it was observed that it could not perform satisfactorily on Indian data (details are captured in Table 1), as the pre-trained models were not trained on such data.

The proposed work was designed to improve Neuro NER capabilities in the following aspects:

- Improve applicability of the system to Indian data
- Improve the efficiency of a de-identification system with NeuroNER

2.0 ORGANIZATION OF PAPER

Section 3 covers the state of the art literature and gives an analysis of the gaps in the existing research which has laid the foundations of the proposed methodology.

Section 4 (Background) describes the Deep Learning model called NeuroNER which is the basis to identify entities of interest in a text.

Section 5 (Proposed Solution) details the analysis of the NeuroNER model and the solution to enhance the model by transfer learning and others techniques, to improve recognition of Indian PHI in a text.

Section 6 (Results and Comparative Analysis with existing techniques) explains the results obtained from the proposed methodologies with performance comparisons to existing techniques.

Section 7 (Discussion and analysis) covers the key findings and result analysis with the existing techniques and proves the hypothesis with a recap of the final outcome.

Section 8 (Conclusion) summarizes and provides insights to the usefulness and application of the proposed solution to relevant areas. This section also provides future directions to build on and improve.

3.0 STATE OF THE ART

In literature, researchers usually follow three standard methods for automated PHI de-identification. They are Rule based, Machine Learning based and Hybrid methods [23].

Rule based de-identification systems [9] are based on extensive hand-coded rules and specialized dictionaries. Rule based systems do not require a large amount of training data but different variations have to be captured. Curating rules requires significant manual work. Rule creators make assumptions on the data, thereby limiting flexibility on unseen data.

Machine learning based de-identification systems try to solve the problem by token classification. In literature, different machine learning algorithms, including CRFs[3] and Support Vector Machines (SVM) [13, 24] have been used. In general, ML-based systems perform better than rule-based systems due to the inherent flexibility. ML based systems perform poorly on PHI types on which limited data is available.

Hybrid systems can combine the benefits of both rules and machine learning [23]. Certain PHI types like dates are best captured using regular expressions whereas PHI types like names are best captured using machine learning techniques. In [15], a hybrid system combining a token-level CRF, a character-level CRF, and a rule-based classifier was used for de-identification.

In recent years there is a noticeable trend in using hybrid methods which combine rule based system and deep learning networks for de-identification tasks [22]. Among the deep learning network architectures, Bi-directional Long Short-Term Memory Networks have been successfully used in the field of Named Entity Recognition [12]. Transfer learning with NeuroNER has been shown to be beneficial for a target set with a small number of labels. To the best of our knowledge, we could not find any publication that exclusively covers Indian PHI de-identification.

4.0 BACKGROUND

Named Entity Recognition, is a technique to identify entities of interest in the text, such as locations, organizations and temporal expressions. The identified entities are used later by applications such as de-identification or information extraction. Also, machine learning systems could use these identified entities for natural language processing tasks [6,17].

The main objective is to identify noun phrases or part of noun phrases automatically from the text. Named entities are usually not simple separate words, but segments or chunks of text. Thus, a parsing prediction model is required to predict whether a group of tokens belongs to the same entity [14].

The main components of NeuroNER are recurrent neural networks (RNNs), in particular, a type of RNN called Long Short Term Memory (LSTM).

The system is composed of three layers [3,5,6,9,11,14,18,19]:

1. The character-enhanced token embedding layer generates a vector representation for each token.

2. The label prediction layer uses the sequence of vector representations corresponding to a sequence of tokens to calculate the probability of each label per corresponding token.

3. Finally, the sequence optimization layer generates the most probable sequence of predicted labels formed on the sequence of probability vectors from the label prediction layer.

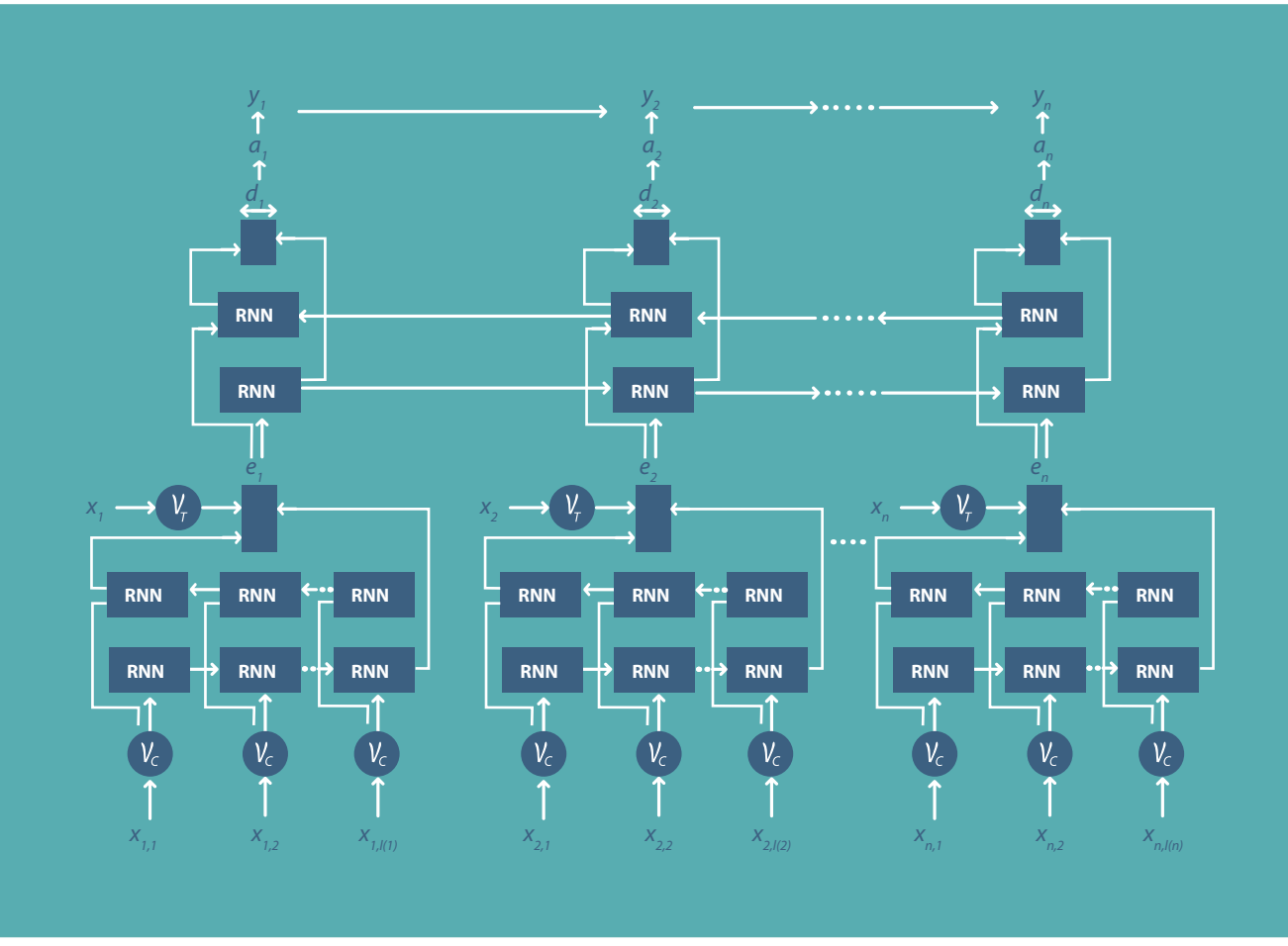


FIGURE 1. ***Architecture of NeuroNER model***

Figure 1 shows the architecture of the NeuroNER neural network. The type of RNN used in this model is Long Short Term Memory (LSTM) [5]

n is the number of tokens

x_i is the i^{th} token.

V_T is the mapping from tokens to token embedding.

$l_{(i)}$ is the number of characters

$x_{i,j}$ is the j^{th} character in the i^{th} token.

V_C is the mapping from characters to character embedding.

e_i is the character-enhanced token embedding of the i^{th} token.

$\overleftrightarrow{d}_i$ is the output of the LSTM of label prediction layer

a^i is the probability vector over labels

y^i is the predicted label of the i^{th} token

5.0 PROPOSED SOLUTION

During the testing with five NeuroNER models, some of the Indian data such as names and locations, were not identified by NeuroNER system satisfactorily. After careful analysis, we proposed to add an additional model trained on Indian data to improve the performance of the NeuroNER to address Indian data.

5.1. Analysis of NeuroNER

NeuroNERs five pre-trained models were used to analyse and test simulated reports with Indian PHI data. During testing, four models, that is, i^2b^2 spaCy, i^2b^2 Stanford, MIMIC spaCy and MIMIC Stanford showed better results compared to CoNLL since it identified many medical terms as PHI. To maintain efficiency, we excluded CoNLL model and used the rest of the four models mentioned above for further experimentation. Figure 2 covers the overall flow of the pre-trained models validation of NeuroNER.

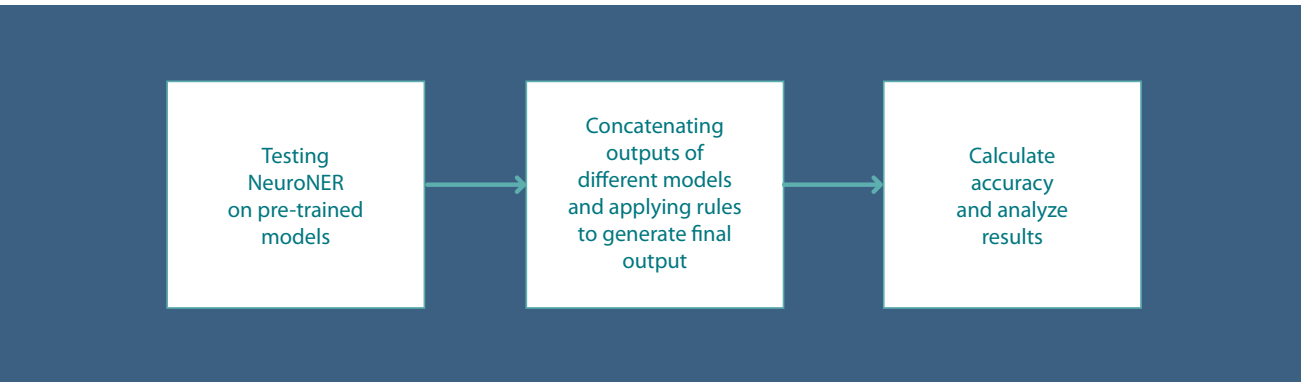


FIGURE 2. ***Testing NeuroNER models***

5.2. Simulation of reports with Indian context

MIMIC-III Data Set: The MIMIC-III dataset contains data for 61,532 ICU stays over 58,976 hospital admissions for 46,520 patients, including 2 million patient notes. We used the MIMIC de-identified medical reports from the MIMIC database. There are various tables present in this database, among which we extracted 5000 medical reports from the NOTEEVENTS table. PostgreSQL was used to load and extract data.

Data Extraction: To make a collection of relevant Indian data, various websites were scraped, with adherence to their privacy rules, to extract Indian names, Indian locations (all towns and cities), Indian Hospital and unit names. Web scraping is a software technique used for extracting information from websites. We used BeautifulSoup, a python software for web scraping data from HTML and XML files. A parse tree was created for parsed pages that can be used to extract data from HTML.

Data Transformation: After scraping websites, the generated data was thoroughly analyzed and their shortcomings were addressed. Various transformations were performed on the data, which included data cleaning, data formatting and transforming data into a suitable form for experimentation.

Data Set Creation: Figure 3 shows the process followed to simulate medical reports with an Indian context. Data extracted through web scrapping stored as lists were used to replace the anonymized name and other de-identified PHI information that exist in the reports extracted from the MIMIC database. It was also made sure that the

simulated reports with Indian context adhere to the format that is suitable for transfer learning with the existing NeuroNER models. A total of 3300 reports were simulated. Out of all the simulated reports, 2000 reports were used for training, 1000 reports were used for test and 300 reports were used for unseen validation.

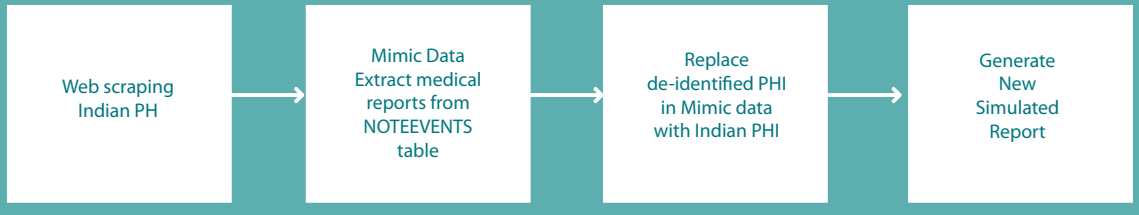


FIGURE 3. ***Process to simulate data set with Indian context***

5.3. Transfer Learning

For transfer learning with Indian context data, we used NeuroNER's MIMIC spaCy pre trained models and performed transfer learning [1, 21]. During the training process, various hyper-parameters such as character based token embedding, LSTM dimension, dropout probability and maximum number of epochs were considered. After the transfer learning, we selected the model with the best epoch (epoch 6 was selected out of 30 epochs) and prepared that model to generate labels in the de-identification process. During this process, we have also introduced new labels which are not standard labels from NeuroNER. To assess the performance, we computed precision, recall, and F-score. Figure 4 covers the overall flow of the proposed de-identification process.

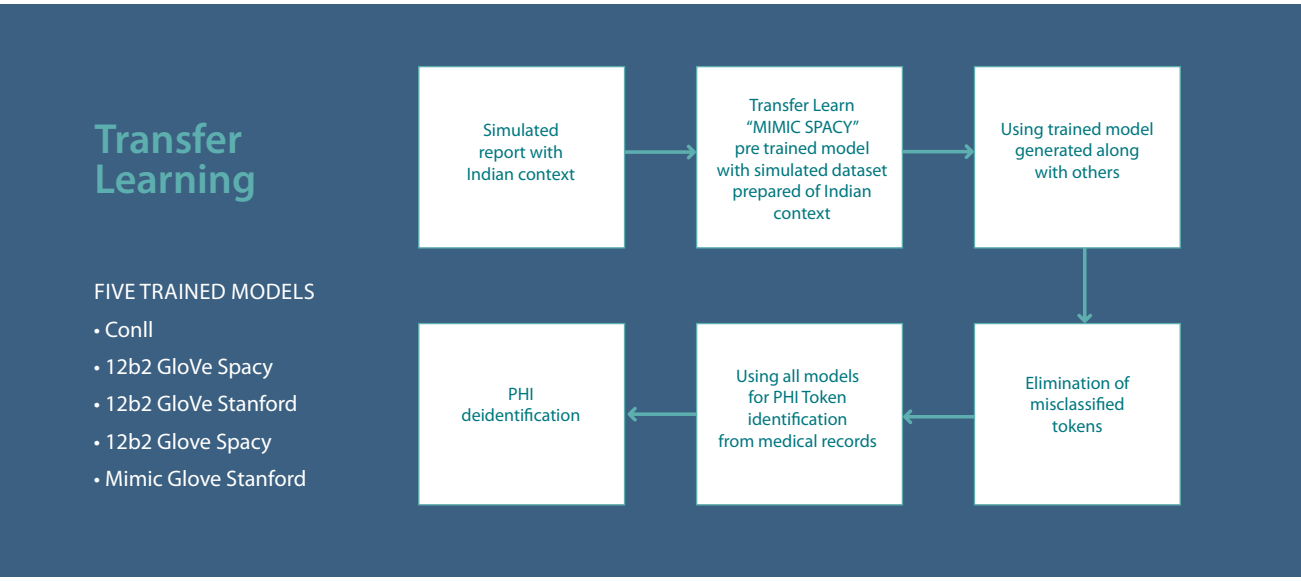


FIGURE 4. ***Proposed de-identification model update***

We calculated precision, recall and F-score as follows:

$$Precision = TP/(TP+FP) \quad (1)$$

$$Recall = TP/(TP+FN) \quad (2)$$

$$F\text{-}Score = (2*Precision*Recall)/(Precision+Recall) \quad (3)$$

TP: True positives, ***FP***: False positives, ***FN***: False negatives

In words, precision is the proportion of the predicted PHI labels that are correct labels. Recall is the proportion of the PHI labels that are predicted correctly and F-score is the harmonic mean of precision and recall. GloVe vectors were used for token embedding.

Eliminate misclassification tokens: For reducing the number of misclassifications, we used a medical dictionary. We scraped medical websites to obtain medical terms and created a hash table to reduce the access time for dictionary lookup. We check for every token, except person and location, if it belongs to any of the medical terms. If so, we label it as "not-PHI" and do not de-identify it.

Concatenation of different model outputs: To improve the efficiency of the system, we developed a technique to concatenate the results from all the pre-trained models. We kept the label-tagging scheme consistent to the MIMIC SpaCy type of labels (12 types). We created a text file where we concatenate outputs of all these models according to specifications. For concatenation we used the weighted averages where the weights were determined by the accuracy of each model for certain PHI information.

PHI de-identification: All the labelled tags of PHI were anonymized in the original reports. Every label was anonymized with a dummy value. Dates were shifted by a few years and seasons were kept intact. Regular expressions were written to mask variables like email, IP address, vehicle number, url and account numbers in case they weren't identified by proposed identifier. After anonymization, de-identified reports are returned as the final output.

6.0 RESULTS AND COMPARATIVE ANALYSIS WITH EXISTING TECHNIQUES

Initially, NeuroNER was tested for its performance on simulated Indian context reports.

Table 2 shows the comparison between the claimed result on foreign data and the result obtained when tested on Indian context medical reports.

	Indian Data			Foreign Data As Reported			
MODEL	PRECISION	RECALL	F-SCORE	MODEL	PRECISION	RECALL	F-SCORE
i2b2 Stanford	78.97	54.81	64.71	i2b2	97.92	97.84	97.88
i2b2 SpaCy	86.86	60.88	71.59				
MIMIC Stanford	97.39	46.09	62.57	MIMIC	98.82	99.40	99.11
MIMIC SpaCy	97.15	52.01	67.75				
CoNLL	19.85	32.78	24.73	CoNLL	Not Available	Not Available	90.50
Miminc SpaCy with transfer learned on Indian data	97.09	97.19	97.14	Not Available			

TABLE 2. ***Comparison of results on different pre-trained models***

Concatenation of the results obtained from the i2b2 Stanford, i2b2 SpaCy, MIMIC Stanford, MIMIC SpaCy and the new MIMIC SpaCy India model, provides reliable results. On the test set, we processed 1816318 tokens with 45889 PHI of which we corrected 44398. Detailed results are captured in Table 3.

PHI Labels	Precision	Recall	F-Score	Frequency of occurrence in test set
AGE	93.24%	60.53%	73.4	74
DATE	99.60%	99.56%	99.58	26536
HOSPITAL	95.40%	93.76%	94.57	4757
IDNUM	99.37%	98.51%	98.94	797
LOCATION OTHER	86.75%	93.68%	90.08	1233
NAME	97.06%	96.51%	96.78	10771
PHONE	99.88%	99.94%	99.91	1721

TABLE 3. ***Results of each PHI label present in the test set***

7.0 DISCUSSION AND ANALYSIS

In this work, 3300 reports were simulated by inserting Indian PHI information in the medical reports extracted from NOTEEVENTS table of the MIMIC database. Out of all the simulated reports, 2000 were used for training, 1000 were used for testing and 300 were used for unseen validation. During the testing of the existing NeuroNER's models on the simulated data, four models, that is, i2b2 spaCy, i2b2 Stanford, MIMIC spaCy and MIMIC Stanford showed better results compared to CoNLL since the latter identified many medical terms as PHI.

To maintain efficiency, we excluded CoNLL model and used the rest of the four models mentioned above for further experimentation. For transfer learning with Indian context data, we used NeuroNER's MIMIC spaCy pre trained model and performed transfer learning. With the transfer learning, the MIMIC SpaCy model on simulated Indian PHI data shows an improvement in F-score from 67.75 to 97.14. Results obtained from concatenation of NeruoNER's four models and the model trained on Indian simulated data provides significant improvement in the processed tokens as shown in Table 3.

8.0 CONCLUSION

To improve the performance of text de-identification on Indian PHI data, NeuroNER's Deep Learning pre-trained models were updated with transfer learning on simulated data. Since the Indian community is spread across the world, the proposed approach can be extended to different English speaking countries to de-identify medical reports. More general, in Deep Learning, more data is better and updating models with more data is a good way to improve PHI de-identification performance. For further improvements, different Deep Learning architectures can be explored.

REFERENCES

1. *Andrew Arnold, Ramesh Nallapati and William W. Cohen. Exploiting Feature Hierarchy for Transfer Learning in Named Entity Recognition". Proceedings of ACL-08: HLT, 2008*
2. *Bin He, Yi Guan, Jianyi Cheng, Keting Cen, and Wenlan Hua. 2015. "Crfs based de-identification of medical records." Journal of biomedical informatics 58:S39–S46.*
3. *Bui, D. D. A., M. Wyatt, and J. J. Cimino, "The UAB informatics institute and 2016 CEGS N-GRID de-identification shared task challenge", Journal of Biomedical Informatics, 2017.*
4. *Dorr DA, et al: "Assessing the difficulty and time cost of de-identification in clinical narratives". Methods Inf Med 2006, 246-52. [5] Franck Dernoncourt, Ji Young Lee, Peter Szolovits, Ozlem Uzuner. "De-identification of Patient Notes with Recurrent Neural Networks." arXiv:1606.03475v1 [cs.CL] 10 Jun 2016*
5. *Franck Dernoncourt, Ji Young Lee, Peter Szolovits. "NeuroNER: an easy-to-use program for named-entity recognition based on neural networks." arXiv:1705.05487 [cs.CL] 16 May 2017*
6. *GPO, U.S: 45 C.F.R. § 46 Protection of Human Subjects 2008 [http://www. access.gpo. gov/nara/cfr/waisidx_08/45cfr46_08.html]*
7. *GPO, U.S: 45 C.F.R. § 164 Security and Privacy 2008 [http://www.access.gpo.gov/ nara/cfr/waisidx_08/45cfr164_08.html].*
8. *Ishna Neamatullah, Margaret M Douglass, Li-wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. "Automated de-identification of free-text medical records." BMC Medical Informatics and Decision Making, 8 (1) (2008), pp. 641-717, 2008 8:32, PMC2526997 2008*
9. *Ji Young Lee, Franck Dernoncourt, Peter Szolovits, "Transfer Learning for Named-Entity Recognition with Neural Networks" arXiv:1705.06273 [cs.CL]*
10. *Ji Young Lee, Franck Dernoncourt, O¨zlem Uzuner, Peter Szolovits. "Feature Augmented Neural Networks for Patient Note De-identification." arXiv: 1610.09704 [cs.CL] 30 Oct 2016*
11. *Kaung Khin, Philipp Burckhardt, Rema Padman, "A Deep Learning Architecture for De-identification of Patient" Notes: Implementation and Evaluation (arXiv:1810.01570 [cs.CL])*

12. *Khalifa, A. and S. Meystre, "Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes", Journal of Biomedical Informatics 58 (Supplement), S128-S132, 2015.*

13. *Lev Ratinov and Dan Roth. "Design Challenges and Misconceptions in Named Entity Recognition." CoNLL '09 Proceedings of the 13th Conference on Computational Natural Language Learning, Stroudsburg, PA, 2009, pp. 147-155.*

14. *Liu, Z., B. Tang, X. Wang, and Q. Chen "De-identification of clinical notes via recurrent neural network and conditional random field" Journal of Biomedical Informatics 75, S34- S42, 2017.*

15. *Meystre et al.: "Automatic de-identification of textual documents in the electronic health record: a review of recent research". BMC Medical Research Methodology 2010 10:70*

16. *Morrison FP, et al: "Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes?" J Am Med Inform Assoc 2009, 16(1):37-9*

17. *Özlem Uzuner, Yuan Luo, Peter Szolovits; "Evaluating the State-of-the-Art in Automatic De-identification." Journal of the American Medical Informatics Association, Volume 14, Issue 5, 1 September 2007, Pages 550–563*

18. *Shweta, Asif Ekbal, Sriparna Saha ,Pushpak Bhattacharyya. "Deep Learning Architecture for Patient Data De-identification in Clinical Records." ClinicalNLP@ COLING 2016*

19. *Szarvas G, Farkas R, Kocsor A: "A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms". 9th Int Conf Disc Sci (DS2006), LNAI 2006, 267-278*

20. *Szarvas G, Farkas R, Busa-Fekete R: "State-of-the-art anonymization of medical records using an iterative machine learning framework". J Am Med Inform Assoc 2007, 574-80*

21. *Vithya Yogarajan, Michael Mayo, "A survey of automatic de-identification of longitudinal clinical narratives", Bernhard Pfahringer "arXiv:1810.06765 [cs.AI]"*

22. *Hee-Jin Lee, Yonghui Wu, Yaoyun Zhang, Jun Xu, Hua Xu, Kirk Roberts, "A hybrid approach to automatic de-identification of psychiatric notes". Journal of Biomedical Informatics 75 (2017) S19–S27.*

23. *Nandkishor P. Karlekar, N Gomathi. "OW-SVM: Ontology and whale optimization based support vector machine for privacy preserved medical data classification in cloud", International Journal of Communication Systems, 2018.*

CHAPTER 4

AUTHORIZATION FRAMEWORK FOR MEDICAL DATA

Geetha Madadevaiah[1]
RV Prasad[1]
Amogh Hiremath[1]
Michel Dumontier[2]
Andre Dekker[3]

International Journal of Database Management Systems (IJDMS)
VOL.11, NO.2/3, JUNE 2019

ABSTRACT

In this paper, the authors describe an approach for sharing sensitive medical data with the consent of the data owner. The framework builds on the advantages of Semantic Web technologies and makes it secure and robust for sharing sensitive information in a controlled environment. The framework uses a combination of Role-Based and Rule-Based Access Policies to provide security to a medical data repository as per the FAIR guidelines. A lightweight ontology was developed, to collect consent from the users indicating which part of their data they want to share with another user having a particular role. Here, the authors have considered the scenario of sharing the medical data by the owner of data, say the patient, with relevant persons such as physicians, researchers, pharmacists, etc. To prove this concept, the authors developed a prototype and validated using the Sesame OpenRDF Workbench with 202,908 triples and a consent graph stating consents per patient.

Keywords

ACCESS POLICIES
SEMANTIC WEB
RDF/SPARQL
ROLE BASED
RULE BASED
FAIR
CONSENT

1.0 INTRODUCTION

Artificial intelligence based learning healthcare systems aim to use information technology and data infrastructures to rapidly apply scientific insights to clinical care and to power scientific discovery from clinical insights [1,2].For this to work, large amounts of routine health care and scientific data need to be made FAIR - Findable, Accessible, Interoperable and Reusable [3] - for both humans and machines. The definition of FAIR principles is specified by the GO FAIR initiative [30].

• **Findable**: Data that is easy to find and identify for computers and humans. It includes metadata to facilitate searching for specific datasets.

• **Accessible**: Data is easily retrievable on long-term data storage. It has well defined access protocols.

• **Interoperable**: Data can be used in combination with other datasets without loss in meaning of terms and values.

• **Reusable**: Data that can be further processed for future research. Adequate information about the data procurement, consent agreements, processing methods (provenance) and license terms are necessary.

However, access to clinical data is a longstanding challenge, particularly given the administrative, political and ethical barriers [4].

Ethical challenges often center on the need to protect the privacy of patients. One ethically and legally accepted way of accessing and processing personal data, such as patient data, is to seek consent of the data subject involved [5]. Such consent must be specific. In the health care context, this means that patients should be able to control access to specific data elements to specific persons or machines for specific uses.

Consent also needs to be dynamic as new data elements become available all the time and patients may change their minds and have a right to be forgotten in some jurisdictions [5]. On top of patient specific control, access policies may also be informed by institutional or national guidelines, like the USA HIPAA law [6], which defines specific data elements to be removed when data is shared with parties not involved in the direct care

of patients, such as scientists. The Automatable Discovery and Access Matrix(ADA-M) is an example of an initiative to encode such guidelines for consumption by machines [7]. Given these requirements, consent systems to get access to clinical data require a finely grained, multi-level (patient, institutional, etc.) and dynamic combination of authentication (which human or machine agent is accessing the data) and authorization (what data is accessed by which agent and for what reason).

Semantic Web technologies can be used to make data FAIR [8] and have been used as first implementations of a rapid learning health care system [9]. The Semantic Web, and its associated standards such as the RDF universal data model [10,11, 28], the SPARQL query language [12,13,25], and the OWL ontology language [14,15], aims to extend the Web from only consisting of human readable documents to a Web of machine understandable data [16]. However, the vision of the Semantic Web is mostly centered on open, linked data [17] so authentication and especially authorization have received relative modest attention.

The aim of this research is to develop a specific consent-based authorization scheme for clinical data using Semantic Web technology. The authors approached this by reviewing state of the art existing methods and techniques and comparing these with the proposed method.

2.0 RELATED WORK

Previous efforts in this domain include the work by Finin et al. [18] who represented a role-based control model into OWL. Their research focuses on the description of role concepts and their relationships (such as hierarchies). However, these investigators did not apply the roles in subsequent access on the Semantic Web. The author Büyükkılıç [24] discusses a framework to develop a semantic IS standard based on a business rule approach. The abstract concept of a rule based approach has been embraced by the authors in the proposed method.

Rishi Kanth Saripalle et al.[27], have addressed the security and privacy at the knowledge level. They have proposed a Role Based Access Control Model to provide permissions to a RDF knowledge source. However, this approach would require administration and maintenance of roles and permissions by a knowledge/database administrator.

Gabillon and Letouzey [19] proposes to create security views on an RDF graph, similar to views managed by relational database administrators. In this approach, a graph is first created (using a SPARQL CONSTRUCT query) in which the security or access policy is applied. Subsequently, this graph is offered to the authorized user. They use a query based enforcement framework where each user specifies a rule for his existing RDF graph, which defines who can view the graph and what part of the graph can be accessed.

The drawback of their approach is that the query framework is cumbersome. It needs to be re-constructed for a new or modified access, requiring maintenance by the graph administrator. This approach is not efficient for the use case wherein the patient maintains their own consent on data distributed across institutions and provides consent at multiple levels, eg. consent for access of complete graph or consent for a node. The authors have built on the work of Gabillon et al. while addressing their limitations.

Sacco and Passant [20] present a lightweight ontology, the Privacy Preference Ontology to restrict access to a particular RDF data. The users have to first specify an access space indicating to which part of the data graph and to whom the restrictions have to be applied. Once the access space is defined, the users can specify the fine-grained access policies to the data belonging to a particular access space. Again, the use case of patients maintaining their own consents, the limitations of such policies is that if there is modification in the dataset, almost all the rules have to undergo modification.

K. Mohan and M. Aramudhan [21] have defined access policies using an ontology-based approach, where they consider Object and Data properties for providing a secure access to personal health data stored in the cloud. They have not addressed the workflow of who generates the rules and the approach lacks a role based approach which makes it easier to define the access policies for healthcare data.

In [22] Hannes Muhleisen et al. define PeLDS (Policy enabled Linked Data Server), where they partition the dataset into different named graphs and create temporary view on those graphs by defining rules. The rules are defined using SWRL (Semantic Web Rule Language). Each rule from the access policy is attributed with an additional consequence to add the rule identifier to a global list of matched rules. If such a rule matches due to sufficient access rights for the current user, it will be added to this list. The list of rules is evaluated, and for every triple matching the data classifications

contained in the rules consequence predicate list is copied from the dataset to the result graph. Rules depend on the structure of the dataset where a view of the graph is obtained and in case the structure is changed, all the rules have to be modified accordingly. The framework can further be extended to other types of data such as DICOM data converted into RDF format [23, 29].

The proposed framework addresses the shortcomings outlined in the above paragraphs. To validate this framework, a prototype of the access policy framework was developed. Institutions and organizations can use this, to share the sensitive RDF data, to specific persons having specific roles. The framework builds on the advantages of the Semantic Web technologies and makes it secure and robust for sharing sensitive information in a controlled environment, as per the Accessible and Re-usable guidelines of FAIR data sharing principles.

3.0 METHODS

Consent Ontology and Graph

A consent ontology was analyzed by Gabillon [19] with which the consents of the user can be collected and stored, see Figure 1. The classes in the ontology are

- **Informed Consent Ontology** (Consent Ontology): This is the main class that defines the framework for collection and storage of consents.

- **Role**: This class specifies the role of the user requesting the data and the user from whom the data is requested.

- **Access Policy**: This class defines the Access Policy defined by the users who would like share their data in a restricted way.

- **Action**: Refers to the type of Action that a user can be perform while requesting the data. The Action may be SPARQL QUERY or a SPARQL UPDATE.

- **Consent**: This class defines the consents under an access policy defined by the user.

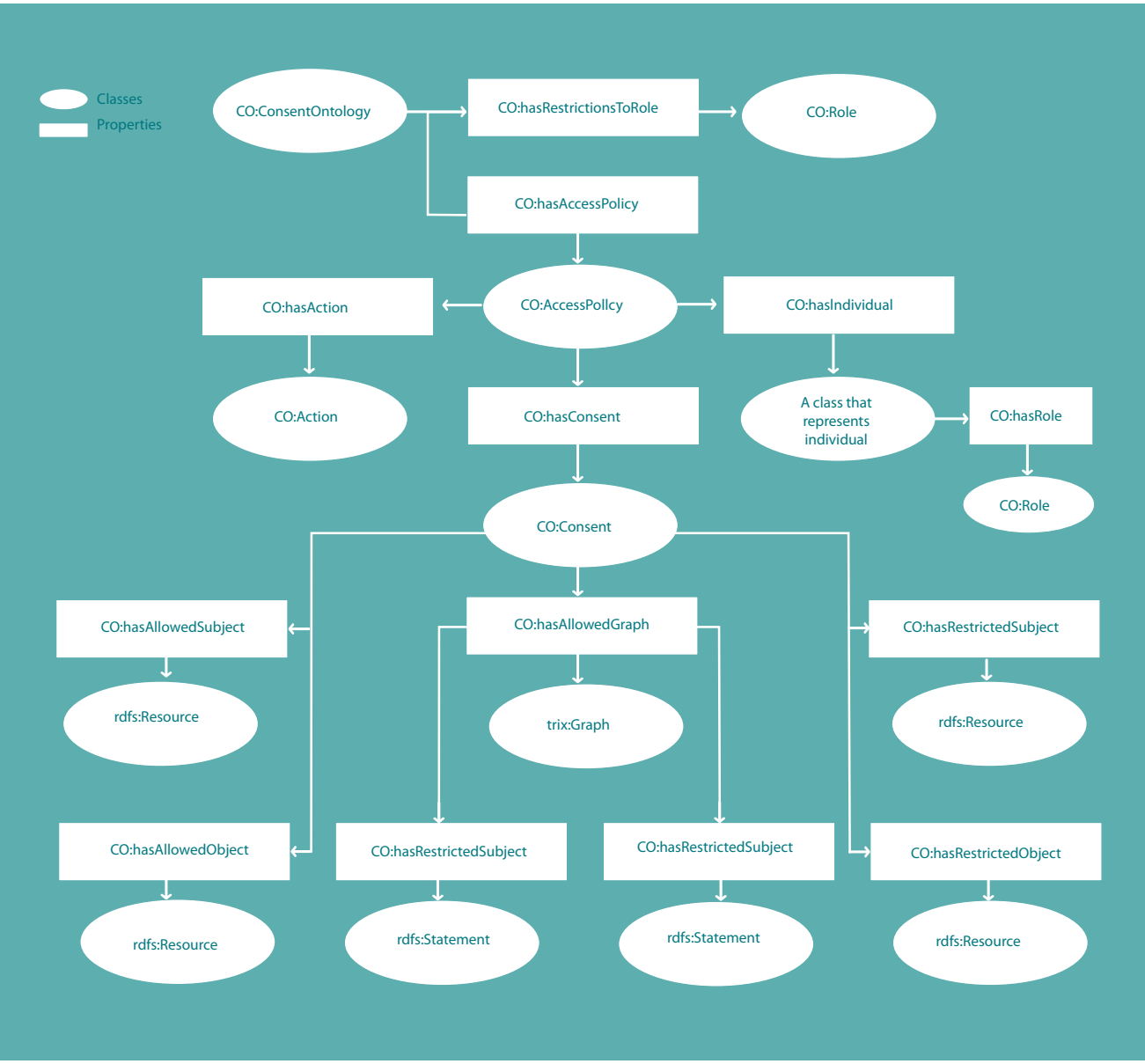


FIGURE 1. ***The consent ontology used in the access policy framework***

CO:ConsentOntology1 a CO:ConsentOntology ;
CO:hasRestrictionsToRole CO:researcherRole.
CO:Access_Policy_1 a CO:AccessPolicy ;
CO:hasAction CO:Query ;
CO:hasConsent CO:Consent_1 ;
CO:hasIndividual <http://www.example.org/1> ;

```
CO:hasRole CO:patientRole .
CO:Consent_1 a CO:Consent ;
CO:hasAllowedSubject <http://www.example.org/icu_data_1> .
CO:researcherRole a CO:Role ;
CO:hasAccess CO:Access_Policy_1.
```

FIGURE 2. ***Access Policy Framework***

The consent ontology was used to define the consent graph for a specific patient. First, roles were defined as instances of the class 'Role' rather than as subclasses, as is recommended by Finin et al. [18]. Then an access policy is defined which includes the consents specified by a user for a given role. The specific consent specified in the example below allows a researcher access to ICU data.

Access Policy Framework

An access policy framework was developed to consume the above described consents and allow users access to the data given their roles (Figure 2). When a user (e.g. a researcher) logs in and requests data from a given patient or patient cohort, a SPARQL CONSTRUCT query is created which creates a subgraph of the original data graph. This subgraph is then shared with the researcher for querying. More specifically, the CONSTRUCT query is a federated SPARQL query across the data graph and the consent graph with a filter for the role of the user. We first query the consent graph with graph identifier consentNamedGraph to obtain the consents. With the obtained consents from the consentNamedGraph, we query the patient graph with graph identifier patientNamedGraph to construct the graph with corresponding triples(Figure 3). Appendix 1 provides a detailed description of the SPARQL query.

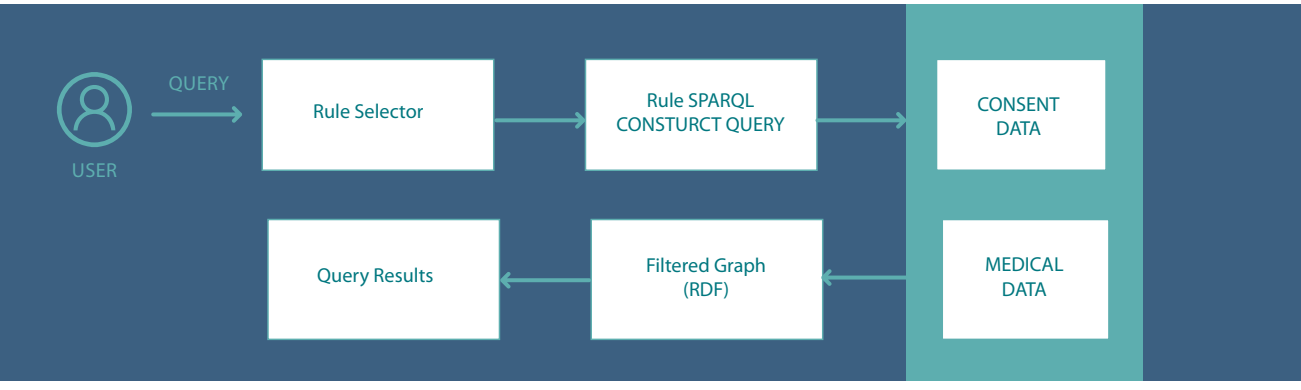


FIGURE 3. ***Generic Rule***

The query as provided in Appendix 1, efficiently retrieves the allowed graph or a combination of allowed subjects and/or allowed properties for a single patient. In order to retrieve allowed graphs for multiple patients from a triple-store, multiple queries need to be constructed using each patient's rules and queried from the triple-store [25].

Before applying the access policies, the data of multiple patients is not stored in a single RDF graph. To enable retrieving data from multiple patients, the complete data is partitioned into different datasets using named graphs. In this way, a graph identifier gets assigned to each triple, thus allowing easy retrieval of all triples belonging to a specific graph. Each dataset with a named graph has an owner who manages the data. In this case, each patient owns a RDF named graph.

4.0 EVALUATION

The access policy framework was evaluated for the following use case :

- A user with the role "Researcher" request the cohort of patients who have consented to the use of a combination of subjects such as EMR data, breast cancer data, ICU data, genetic data or the entire data. .

To evaluate if the access policy framework can fulfill this use cases a number of public datasets containing ICU, HIV and breast cancer data were used, (see Appendix 2 and supplement section at the end of the chapter). Using these, 101 distinct patient graphs with 202,908 triples were constructed.

To specify the consents of all the patients, random subjects were selected and/or properties and/or objects or the entire graph to be allowed by each patient and constructed the consent graph (815 triples, see Appendix 2 and Supplement section below).

All data triples (n~200000) were loaded into Sesame OpenRDF-Workbench (version 4.1.2, Eclipse Foundation) on a Tomcat server (version 7.0, Apache Software Foundation). The evaluation was conducted on a computer with a commodity processor (i5-4310-2GHz, Intel) and 8 GB of RAM. In each case, the SPARQL CONSTRUCT query consuming the consent ontology was executed first to obtain the filtered subgraph for each of the patients.

The procedure is as follows :

- The consent graph and patient graphs are loaded into the triple-store.

- Then a single query and retrieval is done for each patient graph leading to same number of queries as the number of patient graphs.

5.0 RESULTS

For the purpose of demonstrating our Access Policy Framework, the authors have considered a RDF graph with a patient graph identifier as "ex:graph_patient1" and consent graph identifier as "ex:graph_consent" using synthetic medical data. Figure 4 and Figure 5 show a RDF document of a patient with ICU acquired data, EMR data, genetic information, breast cancer data and the Consent Ontology.

```
ex:patient1 a ex:Patient;
  ex:has_breastcancer _data ex:breastcancer_data_patient1;
  ex:has_emr_data ex:emr_data_patient1;
  ex:has_genetic_data ex:genetic_data_patient1;
  ex:has_icu_data ex:icu_data_patient1.

ex:breastcancer_data_patient1 a BREASTCANCER:Data;
  BREASTCANCER:hasDocument ex:breastcancer_document_patient1.

ex:emr_data_patient1 a EMR:Data;
  EMR:hasDocument ex:emr_document_patient1.

ex:genetic_data_patient1 a GENETIC:Data;
  GENETIC:hasOocument ex:genetic_document_patient1.

ex:icu_data_patient1 a ICU:Data;
  ICU:hasDocument ex:icu_document_patient1.

ex:breastcancer_document_patient1 a BREASTCANCER:DOCUMENT;
  BREASTCANCER:Age "66"^^xsd:string;
  BREASTCANCER:PR-Status "Positive"^^xsd:string;
  BREASTCANCER:Tumor "Tl"^^xsd:string.

ex:emr_document_patient1 a EMR:DOCUMENT;
  EMR:cost "211.66"^^xsd:string;
  EMR:diagnosis "ENCEFALOPATIA NAO ESPECIFICADA"^^xsd:string;
  EMR:responsible_physician IIAA26E70"^^xsd:string.

ex:genetic_document_patient1 a GENETIC:DOCUMENT;
  GENETIC:PR-Seq "CCTCAAATCACTCI I IGGCAAC"^^xsd:string;
  GENETIC:Patient1D "8"^^xsd:string;
  GENETIC:RT-Seq "CCCATAAGTCCTATTGAAACTGTACC"^^xsd :string.

ex:icu_document_patient1 a ICU:OOCUMENT;
  ICU:Heart_Rate_value 61;
  ICU:Respiratory_Rate_value 10.
```

FIGURE 4. ***RDF Document : ICU, EMR, genetic and breast cancer data***

```
CO:ConsentOntology1      a                        CO:ConsentOntology.
CO:ConsentOntology1      CO:hasRestrictionsTo     CO:researcherRole.

CO:AccessPolicy1         a                        CO:AccessPolicy.
CO:AccessPolicy1         CO:haslndividual         ex:patient1.
CO:AccessPolicy1         CO:hasRole               CO:patientRole.

CO:consent1              a                        CO:Consent.
CO:AccessPolicy1         CO:hasConsent            CO:consent1.

CO:consent1              CO:hasAllowedSubject     ex:emr_data_patient1
```

FIGURE 5. ***Consent Ontology***

The authors consider a scenario in which the medical data of a patient in a hospital has to be shared with a researcher. The authors evaluated the proposed access policies for three different cases in which the consents of the patient given to a researcher are different. For each of the cases, the rules applied were the same as described in the METHODS section.

Case 1: Patient shares only the EMR data

The consent graph for this case is shown in Figure 6, where the patient has allowed the subject ex:emr_data_patient1, which corresponds to the EMR data of the patient.

```
ex:emr_data_patient1 a EMR:Data;
  EMR:hasDocument ex:emr_document_patient1.


ex:emr_document_patient1 a EMR:DOCUMENT;
  EMR:cost “211.66”^^xsd:string;
  EMR:diagnosis ”ENCEFALOPATIA NAOESPE1CIFICADA’’^^xsd:string;
  EMR:responsible_physician “AA26E70”^^xsd:string.
```

FIGURE 6. ***Consent Graph EMR Data***

Figure 7 shows the filtered graph obtained after the rule is applied on the consents and the patient data. The filtered graph is a sub graph of the patient data graph starting from the subject specified in the consents.

```
CO:ConsentOntology1      a                        CO:ConsentOntology.
CO:ConsentOntology1      CO:hasRestrictionsTo     CO:researcherRole.

CO:AccessPolicy1         a                        CO:AccessPolicy.
CO:AccessPolicy1         CO:haslndividual         ex:patient1.
CO:AccessPolicy1         CO:hasRole               CO:patientRole.

CO:consent1              a                        CO:Consent.
CO:AccessPolicy1         CO:hasConsent            CO:consent1.

CO:consent1              CO:hasAllowedProperty    ex:Heart_Rate_value
CO:consent1              CO:hasAllowedSubject     ex:breastcancer_data_patient
```

FIGURE 7. ***Filtered Graph post application of consents***

Case 2: Patients shares only heart rate in the ICU data and the Breast cancer data shown in Figure 8.

```
ex:icu_document_patient1 a ICU:DOCUMENT;
  ICU::Heart_Rate_value 61;

ex:breastcancer_data_patient1 a BREASTCANCER:Data;
  BREASTCANCER:hasDocument ex:breastcancer_document_patient1.

ex:breastcancer_document_patient1 a BREASTCANCER:DOCUMENT;
  BREASTCANCER:Age"66"^^xsd:string;
  BREASTCANCER:PR-Status"Positive"^^xsd:string;
  BREASTCANCER:Tumor "T1"^^xsd:string.
```

FIGURE 8. ***Heart rate in ICU Data and Breast Cancer Data***

Case 3: Patient shares the complete data

Figure 9 shows the consent given by the patient to share the complete graph. There the complete graph with the specific graph identifier will be constructed as a result and shared with the researcher.

```
CO:ConsentOntology1      a                        CO:ConsentOntology.
CO:ConsentOntology1      CO:hasRestrictionsTo     CO:researcherRole.

CO:AccessPolicy1         a                        CO:AccessPolicy.
CO:AccessPolicy1         CO:haslndividual         ex:patient1.
CO:AccessPolicy1         CO:hasRole               CO:patientRole.

CO:consent1              a                        CO:Consent.
CO:AccessPolicy1         CO:hasConsent            CO:consent1.

CO:consent1              CO:hasAllowedGraph       ex:graph_patient
```

FIGURE 9. ***Patient Consent to share full graph***

The total query execution time for all the patients was 111.64 seconds with 70,190 triples constructed as a part of the subgraph or the filtered graph. The total query execution time refers to getting all the filtered graphs for the patient in the triple store. The measurements show a significant improvement over standard SQL based query and retrieve methods.

6.0 DISCUSSION

The framework uses a combination of Role Based and Rule Based Access Policies to provide security to a medical data repository. The prototype is validated using Sesame OpenRDF Workbench with 202,908 triples and a consent graph stating consents per patient. The main advantage of this Access Policy is that there is no requirement for each user to specify the rules. The user will only have to provide the consents. The Central Authority or an Administrator who is responsible for the medical database can specify the rules. The rules are specified according to the structure of the data, irrespective of the consents given by the users.

Various authors have presented their work on access policies for the sensitive RDF data. Finin et al. [18] integrated RBAC (Role Based Access Control) model into OWL. They used OWL ontologies to represent a RBAC model which specifies the access control policies. Gabillon and Letouzey [19] emphasize providing access control policies over named graphs and views which generate a subgraph. They use a query based enforcement framework where each user specifies a rule for his existing RDF graph, which defines who can view the graph and what part of the graph can be accessed. Sacco and Passant [20] present a light weight ontology, the Privacy Preference Ontology(PPO) to restrict access to particular RDF data. The users have to first specify an access space indicating to which part of the data graph and to whom the restrictions have to be applied. Once the access space is defined, the users can specify the fine-grained access policies to the data belonging to a particular access space. The limitations of such policies is that if there is modification in the dataset, almost all the rules have to undergo modification.

However, in each of the above approaches, users have to define their own rule which becomes cumbersome and difficult to manage. Also, in organizations like hospitals, if each patient defines their own rules, it leads to duplication of the rules since the structure of the dataset remain the same. Instead in the proposed framework, a single rule can be defined for all the patients with only customizations in consents.

In this study, the use case of a "Researcher" requesting the consent of a patient was successfully tested.

In addition, the authors believe that the framework can be evaluated successfully for the following scenarios too.

- A patient gives consent to the use of her birth data to a user with the role "Physician" but not to a user with the role "Researcher".

- A patient gives consent to a specific named user with the role "Physician" but not to another named user with the role "Physician". (e.g. a patient may have a conflict with a certain physician)

- A patient gives consent but the data holder withholds consent (e.g. a patient may give consent to share a physician name with the outside world, which the hospital does not allow)

Artificial intelligence approaches incl. machine-learning algorithms require training on large sets of curated data. In a typical scenario, to train an algorithm, the researcher would obtain consent from the data owner; protect the privacy (via de-identification [31]) as the initial process before using the data for further analysis. When the same researcher requires the same data for a different purpose, the researcher maybe required to request another consent from the data owner. Currently, the method and techniques for seeking consent and maintaining a record is at best semi-automated and in many cases a manual process. The data owner may not be fully aware of the consent process and its real world deployment. Similarly, the researcher finds it challenging to obtain data, which has the necessary consent approvals and adheres to the country specific privacy and security laws. The re-usability of data per the FAIR guidelines, enabling data usage for future research and further computational processing with satisfactory licenses and provenance, is an area requiring improvement with process, tools and techniques.

This framework is a step in the direction of empowering the data owner to manage the rights of data usage. Extending this concept further, the framework can include audit trails, history, duration for data usage and purpose of usage.

7.0 CONCLUSION

The authors built a lightweight ontology to collect the consents from the users indicating which data entities they want to share with another user with a specific role. To address concerns of security and privacy of medical data, the authors considered the scenario of sharing the medical data of a hospital for research with the consent of the patient. The authors have followed a hybrid approach with a combination of Role-based and Rule-based Access to provide security to medical RDF data. A central authority, knowledgeable in semantic web and the database constructs the rules, and the patients provide the consents. The advantages of the authorization framework is the ease by which a user can change access rights, by modifying only the consents. Similarly, when there is an update to the dataset, only the rules specified by the central authority requires modification.

Acknowledgements

The authors thank Dr. Shyam Vasudev Rao, Technical Director of Maastricht Education and Research Centre (MERC, Pvt Ltd) Bangalore, for his support and valuable guidance. They acknowledge the contributions of colleagues at Philips Research, Bangalore, specially Ravindra Patil and Kiran Kumar Y for their valuable review comments.

REFERENCES

1. *Maddox TM, Albert NM, Borden WB, Curtis LH, Ferguson TB, Kao DP, et al. The Learning Healthcare System and Cardiovascular Care: A Scientific Statement From the American Heart Association. Circulation. 2017;135:e826–57.*
2. *Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, et al. 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. Radiother. Oncol. 2013;109:159–64*
3. *Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data. 2016;3:160018.*
4. *Sullivan R, Peppercorn J, Sikora K, Zalcberg J, Meropol NJ, Amir E, et al. Delivering affordable cancer care in high-income countries. Lancet Oncol. 2011;12:933–80.*
5. *Regulation GDP. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. Off. J. Eur. Union OJ. 2016;59:1–88.*
6. *Standards for privacy of individually identifiable health information. Office of the Assistant Secretary for Planning and Evaluation, DHHS. Final rule. Fed. Regist. 2000;65:82462–829.*
7. *J. Patrick Woolley, Emily Kirby, Josh Leslie, Francis Jeanson, Moran N. Cabili, Gregory Rushton, et al.Responsible sharing of biomedical data and biospecimens via the "Automatable Discovery and Access Matrix" (ADA-M). npj Genomic Medicine volume 3, Article number: 17 July-2018.*
8. *Wilkinson MD, Verborgh R, Bonino da Silva Santos LO, Clark T, Swertz MA, Kelpin FDL, et al. Interoperability and FAIRness through a novel combination of Web technologies. PeerJ Comput. Sci. 2017;3:e110.*

9. *Jochems A, Deist TM, Soest J van, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Radiother. Oncol. 2016;121:459–67.*

10. *RDF Primer [Internet]. [cited 2017 Jun 8]. Available from: https://www.w3.org/TR/2004/REC-rdf-primer-20040210/*

11. *Decker S, Mitra P, Melnik S. Framework for the semantic Web: an RDF tutorial. IEEE Internet Comput. 2000;4:68–73.*

12. *SPARQL Query Language for RDF [Internet]. [cited 2017 Jun 8]. Available from: https://www.w3.org/TR/rdf-sparql-query/*

13. *Arenas M, Pérez J. Querying semantic web data with SPARQL. Proc. Thirtieth ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. [Internet]. ACM; 2011 [cited 2017 Jun 8]. p. 305–316. Available from: http://dl.acm.org/citation.cfm?id=1989312*

14. *OWL 2 Web Ontology Language Primer (Second Edition) [Internet]. [cited 2017 Jun 9]. Available from: https://www.w3.org/TR/owl2-primer/*

15. *Jonquet C, Shah N, Youn C, Callendar C, Storey M-A, Musen M. NCBO annotator: semantic annotation of biomedical data. Int. Semantic Web Conf. Poster Demo Sess. [Internet]. 2009 [cited 2017 Jun 8]. Available from: http://www.lirmm.fr/~jonquet/publications/documents/Demo-ISWC09-Jonquet.pdf*

16. *Berners-Lee T, Hendler J. Publishing on the semantic web. Nature. 2001;410:1023–4.*

17. *Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R, et al. DBpedia-A crystallization point for the Web of Data. Web Semant. Sci. Serv. Agents World Wide Web. 2009;7:154–165*

18. *Finin T, Joshi A, Kagal L, Niu J, Sandhu R, Winsborough W, et al. R OWL BAC: representing role based access control in OWL. Proc. 13th ACM Symp. Access Control Models Tec Symp. Access Control Models Technol. [Internet]. ACM; 2008 [cited 2017 Jun 8]. p. 73–82. Available from: http://dl.acm.org/citation.cfm?id=1377849hnol. [Internet]. ACM; 2008 [cited 2017 Jun 8]. p. 73–82. Available from: http://dl.acm.org/citation.cfm?id=1377849*

19. *Gabillon A, Letouzey L. A View Based Access Control Model for SPARQL. IEEE; 2010 [cited 2017 Jun 8]. p. 105–12. Available from: http://ieeexplore.ieee.org/document/5636084/*

20. *Sacco O, Passant A. A Privacy Preference Ontology (PPO) for Linked Data. LDOW [Internet]. 2011 [cited 2017 Jun 8]. Available from: http://www.academia.edu/download/6230668/ldow2011-paper01.pdf*

21. *Mohan K, Aramudhan M. Ontology based access control model for healthcare system in cloud computing. Indian J. Sci. Technol. 2015;8:218–222.*

22. *Muhleisen H, Kost M, Freytag J-C. SWRL-based access policies for linked data. Procs SPOT [Internet]. 2010 [cited 2017 Jun 8];80. Available from: http://ceur-ws.org/Vol-576/paper1.pdf*

23. *Mahadevaiah G, Soest JV, Dekker A, Udupa N, Rao SV, Kumar YK, et al. Semantic Representation of Radiotherapy data for effective data mining. Proc. Fifth Int. Conf. Adv. Appl. Sci. Environ. Eng. - ASEE 2016 [Internet]. Kuala Lumpur, Malaysia: Institute of Research Engineers and Doctors, USA; [cited 2017 Jun 8]. p. 12–5. Available from: http://www.seekdl.org/nm.php?id=7421*

24. *Büyükkiliç, Thomas. Rule-based semantic standards: a conceptual framework for rule-based semantic IS standards development, University of Twente 2011.*

25. *http://ai.ia.agh.edu.pl/wiki/_media/pl:dydaktyka:semantic_web:sparql.pdf*

26. *Hira Asghar, Zahid Anwar, Khalid Latif: A deliberately insecure RDF-based Semantic Web application framework for teaching SPARQL/SPARUL injection attacks and defense mechanisms. Computers & Security 58: 63-82 (2016)*

27. *Rishi Kanth Saripalle, Alberto De la Rosa Algarin, Timoteus B. Ziminski: Towards knowledge level privacy and security using RDF/RDFS and RBAC. ICSC 2015: 264-267*

28. *Csilla Farkas, Vaibhav Gowadia, Amit Jain, D. Roy: From XML to RDF: Syntax, Semantics, Security, and Integrity (Invited Paper). IICIS 2004: 41-55*

29. *Vaibhav Gowadia, Csilla Farkas: RDF metadata for XML access control. XML Security 2003: 39-48*

30. *Go FAIR Initiative : https://www.go-fair.org/fair-principles/*

31. *Özlem Uzuner, Yuan Luo, Peter Szolovits; "Evaluating the State-of-the-Art in Automatic De-identification." Journal of the American Medical Informatics Association, Volume 14, Issue 5, 1 September 2007, Pages 550–563*

APPENDIX 1

The source code of a query that efficiently retrieves the allowed graph or a combination of allowed subjects and/or allowed properties for a single patient is provided below.

```
PREFIX BREASTCANCER: <http://breastcancer-data.org/>
PREFIX BREASTCANCERDOCUMENT: <http://breastcancer-data.org/
document#>
PREFIX CO: <http://www.semanticweb.org/310204290/
ontologies/2016/3/ConsentOntology#>
PREFIX sedi: <http://semantic-dicom.org/dcm#>
PREFIX EMR: <http://emr-data.org/>
PREFIX DICOM: <http://dicom-data.org/>
PREFIX DICOMDOCUMENT: <http://dicom-data.org/document#>
PREFIX EMRDOCUMENT: <http://emr-data.org/document#>
PREFIX GENETIC: <http://genetic-data.org/>
PREFIX GENETICDOCUMENT: <http://genetic-data.org/document#>
PREFIX GRAPHIDENTIFIER: <http://www.namedGraph.org/>
PREFIX ICU: <http://icu-data.org/>
PREFIX ICUDOCUMENT: <http://icu-data.org/document#>
PREFIX OCKR: <http://www.example.org/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xml: <http://www.w3.org/XML/1998/namespace>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
//Comments: Based on Patient Name graph and Consent Name
graph, constructing the SPARQL Query to define access policy

CONSTRUCT{ ?s ?p ?o. }

FROM NAMED<patientNamedGraph>
FROM NAMED<consentNamedGraph>

WHERE {
  {
    GRAPH<patientNamedGraph>
    {
      VALUES ?s {OCKR:patient_name}
      ?s ?p ?o.
    }
  }
```

```
  UNION
  { // subgraph is shared with the researcher for querying
    GRAPH<consentNamedGraph>
    {
      ?consentontology CO:hasRestrictionsToRole ?role.
      FILTER(?role = CO:researcherRole)
      ?role CO:hasAccess ?accessPolicy.
      ?accessPolicy CO:hasConsent ?consent.
      ?accessPolicy CO:hasIndividual OCKR:patient_name.
    }
    {
      GRAPH<consentNamedGraph>
      {
        ?consent CO:hasAllowedSubject ?s.
      }
      GRAPH<patientNamedGraph>
      {
        ?s ?p ?o.
      }
    }
    UNION
    {    //query the consent graph with graph identifier
consentNamedGraph to obtain the consents
      GRAPH<consentNamedGraph>
      {
        ?consent CO:hasAllowedSubject ?subject.
      }
      GRAPH<patientNamedGraph>
      {
        ?subject (!CO:)+ ?s.
        ?s ?p ?o.
      }
    }
```

```
    UNION
    {
      GRAPH<consentNamedGraph>
      {
        ?consent CO:hasAllowedProperty ?p.
      }
//query the patient graph with graph identifier
patientNamedGraph
      GRAPH<patientNamedGraph>
      {
        ?s ?p ?o.
      }
    }
    UNION
    {      GRAPH<consentNamedGraph>
      {
        ?consent CO:hasAllowedProperty ?property.
      }
      GRAPH<patientNamedGraph>
      {
        ?s ?property ?o1.
        VALUES ?p {rdf:type}
        ?s ?p ?o.
      }
    }
    UNION
    {
      GRAPH<consentNamedGraph>
      {
        ?consent CO:hasAllowedProperty ?property.
      }
        GRAPH<patientNamedGraph>
      {
        ?s1 ?property ?s.
        VALUES ?p {rdf:type}
        ?s ?p ?o.
      }
    }// construct the graph with corresponding triples
    UNION
    {
      GRAPH<consentNamedGraph>
      {
        ?consent CO:hasAllowedGraph ?graph.
        GRAPH ?graph {?s ?p ?o.}
      }
    }
  }
}
```

APPENDIX 2

The dataset of with 202,908 triples patients has be converted into RDF format by collecting data from different sources. The sources and description of the data is given below:

HIV Genetic Data

https://www.kaggle.com/c/hivprogression/data

The Dataset is organized as follows:

Col-1: patient ID

Col-2: responder status ("1" for patients who improved and "0" otherwise)

Col-3: Protease nucleotide sequence (if available)

Col-4: Reverse Transcriptase nucleotide sequence (if available)

Col-5: viral load at the beginning of therapy (log-10 units)

Col-6: CD4 count at the beginning of therapy

The Responder status indicates whether the patient improved after 16 weeks of therapy. Improvement is defined as a 100-fold decrease in the HIV-1 viral load.

There's a brief description of Protease nucleotide sequence, Reverse Transcriptase nucleotide sequence, viral load and CD4 count on the background page.

Breast Cancer Data

https://www.kaggle.com/piotrgrabo/breastcancerproteomes/

Breast Cancer Proteomes

First column "Complete TCGA ID" are the sample IDs. All other columns have self-explanatory names, contain data about the cancer classification of a given sample using different methods.

Mimic ICU Data

https://physionet.org/physiobank/database/mimicdb/

The complete MIMIC Clinical Database is described in [14].

SUPPLEMENT A

BASIS OF TRIPLET PATTERN

Basic Graph Pattern – set of Triple Patterns.

Triple Pattern

Similar to an RDF Triple (subject, predicate, object), but any component can be a query variable; literal subjects are allowed –

```
?book dc:title ?title
```

Matching a Triple Pattern to a Graph

Bindings between variables and RDF Terms

Group Pattern

A set of graph patterns must all match Value Constraints - restrict RDF terms in a solution.

Optional Graph Patterns

Additional patterns may extend the solution.

Matching of Basic Graph Patterns

A Pattern Solution of Graph Pattern GP on graph G is any substitution S such that S(GP) is a subgraph of G

```
SELECT ?x ?v WHERE { ?x ?x ?v }
rdf:type rdf:type rdf:Property
```

Ref: http://ai.ia.agh.edu.pl/wiki/_media/pl:dydaktyka:semantic_web:sparql.pdf

CHAPTER 5

SEMANTIC REPRESENTATION OF RADIOTHERAPY DATA FOR EFFECTIVE DATA MINING

Geetha Mahadevaiah
Johan Van Soest
Dr. Andre Dekker
Dr. Narendranath Udupa
Dr. Shyam Vasudev Rao
Y.Kiran Kumar
R.V.Prasad

International Journal of Biomedical Science & Bioinformatics
VOLUME 3 : ISSUE 2 [ISSN 2475 - 2290] 31 AUGUST 2016

ABSTRACT

Radiotherapy plays an important role in the treatment of cancer patients. As part of the clinical workflow, a patient has to undergo diagnostic imaging procedures, which are used to identify the tumor location and size. Enormous amounts of data are generated during this procedure. The volume of medical information is so large and complex that it becomes difficult to mine for relevant information. The Digital Imaging and Communications in Medicine (DICOM) standard is widely used in medicine for storing and transmitting medical image information. DICOM-RT is an extension to DICOM standard, and dedicated to radiotherapy. In this paper, the researchers propose a technique to extract and store clinically relevant features from DICOM files using semantic concepts. The proposed technique defines a novel method to flatten the hierarchy of DICOM-RT for storing the clinically relevant information into triples in a Resource Description Framework (RDF) repository. The methodology also proposes different combinations for storing data such as DICOM-RT with tumor information and DICOM-RT with pathology details. The proposed method uses Semantic Web Technology to store and represent the information from DICOM-RT files as a RDF graph. Natural Language Processing is used for mining the data. The researchers have evaluated the methodology qualitatively for 20 patients including combinations such as RTSTRUCT, tumor size data along with CT data, pathology information, by producing 25 varieties of different queries. The researchers have used the proposed methodology in a quantitative analysis for different hypothetical conditions resulting in an accuracy of 90%.

Keywords

DICOM-RT
SEMANTIC WEB
RDF
SPARQL
NATURAL LANGUAGE PROCESSING

1.0 INTRODUCTION

Radiation therapy plays an important role in the treatment of cancer patients [1]. In the radiation therapy care process, patients have to undergo imaging procedures, which are performed to identify the tumor location and size. Data generated during this procedure contains large volumes of information as well as complex structures, which makes it a challenging task for clinicians to query and retrieve relevant data [2]. Standards like the Digital Imaging and Communications in Medicine (DICOM) [3] standard is widely used in radiology for diagnostic imaging. The DICOM standard has evolved over the years and extended to incorporate medical specialties such as radiotherapy, which led to the creation of the DICOM radiation therapy standard (DICOM-RT) [4]. DICOM-RT objects provide information about patient related structures / regions of interest identified from imaging known as radiotherapy structure set (RTSTRUCT), radiotherapy treatment plan information (RTPLAN) and the planned radiation dose distributions (RTDOSE) [5]. The DICOM-RT objects are stored in an hierarchical manner and this restricts the search path when traversing various DICOM objects. The DICOM query model and current DICOM tools do not support the required traversing well [6]. But in radiotherapy, traversing objects is necessary as the required information is fragmented across DICOM-RT objects – as described above: The regions of interest for radiotherapy (eg. the tumor and critical organ) are stored in a RTSTRUCT object. The coordinates and slices are defined in the CT objects, the treatment information are stored in the RTPLAN object and the radiation dose matrix in the RTDOSE objects, see Figure 1 [7].

Semantic Web technology provides access to data and enables context based information interpretation of any data source. The graph based data representation enables dynamic modelling, and the use of standardized ontologies allows collaboration, sharing and reuse across applications. In the semantic world, data is modelled using the Resource Description Framework (RDF) [8], where resources and their relationships are stored in the form of a "triple" that is in subject-object-predicate (SOP) structure [9,10].

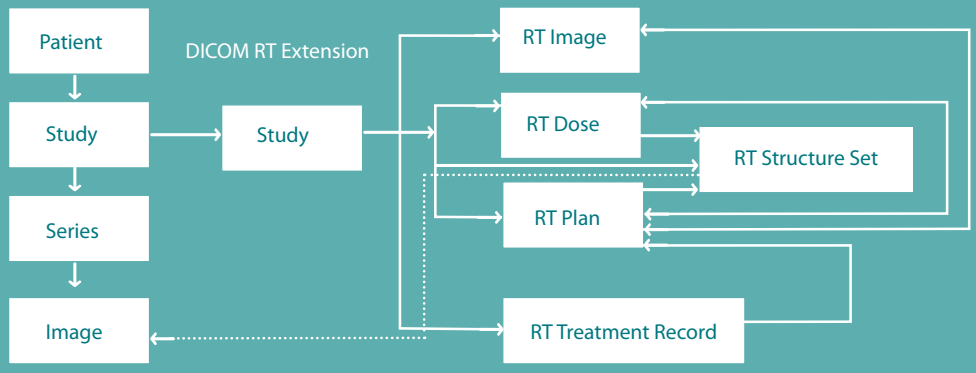


FIGURE 1. ***Illustration of DICOM-RT objects as an extension of the DICOM standard.***

Regarding the use of the Semantic Web in medical imaging, the study by Möller [11] describes semantic techniques for annotating images with concepts using standard ontologies and storing the annotated data in RDF format such that images themselves becomes semantically rich. However, their notion of context is solely tied to anatomy and disease. The research work by Brunnbauer [12] focuses on representing the DICOM metadata by building an ontology. They have developed a tool called "dicom2rdf" for converting the DICOM metadata into RDF. This tool extracts the RDF metadata out of DICOM files and generates RDF files of large size, which includes the pixel data. The present researchers have studied [13] for storing and representing only the metadata of DICOM files in RDF repository. They have created an ontology called SEDI: Semantic DICOM, to represent DICOM metadata elements and developed a proof of concept for storing the DICOM metadata in an RDF repository.

This paper describes a study with the aim to leverage Semantic Web Technology to store the DICOM data for radiotherapy into RDF graphs and a method to mine the data by using a natural language-based search. The retrieval of data from an RDF graph can be achieved by querying using the Simple Protocol and RDF Query Language (SPARQL) [14], which matches the pattern of a statement in the graph. In this paper, the solution to traversing of objects is addressed by representing the objects of DICOM-RT in a generic and flexible format, in contrast to traditional relational databases, where the data is stored in rows and column. The data can also be stored as nodes that are connected to each other using 'semantic' links [15].

2.0 MATERIAL AND METHODS

In this section, we propose a technique based on Semantic Web technology to model metadata of a DICOM-RT object and store it into a RDF repository in the form of triples. This section also explains the process of information retrieval using SPARQL queries and the translation of natural language queries into SPARQL.

2.0 Modeling of DICOM Metadata

The DICOM-RT objects form a hierarchy of data structures, which defines their directionality. The method to extract key information from each of the DICOM-RT objects is crucial. In this study we assume that each DICOM-RT object has "i" number of tags from which the researchers select a set of "j" tags for semantic modelling. The basic necessity to start semantic modelling for RT objects, is to understand the complexity of how data is stored, bring semantics into data, and make data accessible during search. Semantic modelling is similar to conceptual modeling approaches such as the entity-relationship [18]. The ontology is the core part of the semantic system. The ontology provides the domain information in terms of concepts and the relationship between them as properties. RDF is used to make statements about resources relationships. The Figure 2 shows an example of RDF graph creation.

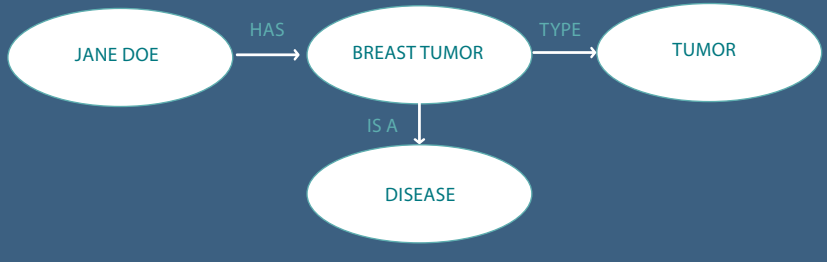


FIGURE 2. ***Shows a RDF graph for "Jane Doe has breast tumor".***

Efforts of the Semantic Web Community made it possible to formalize the knowledge in languages such as Web Ontology Language (OWL) [16]. In this study, the "Semantic DICOM ontology" (SEDI) is used. SEDI is still under development and so far holds the basic structure/concept of DICOM standard. This study has extended the current ontology for a few key RT objects.

Methodology

The RTSTRUCT object contains the information about the region of interest (ROI) in radiation therapy such as organs at risk and tumor volumes (GTV, CTV, and PTV). Each structure set is linked with a frame of reference to the scanned images (usually CT). An RTDOSE objects similarly has a frame of reference to the scanned images but also refers to the RTSTRUCT object – e.g. through the referenced ROI.A part of the hierarchical structure of RT STRUCT and RT DOSE is shown in Figure 3.

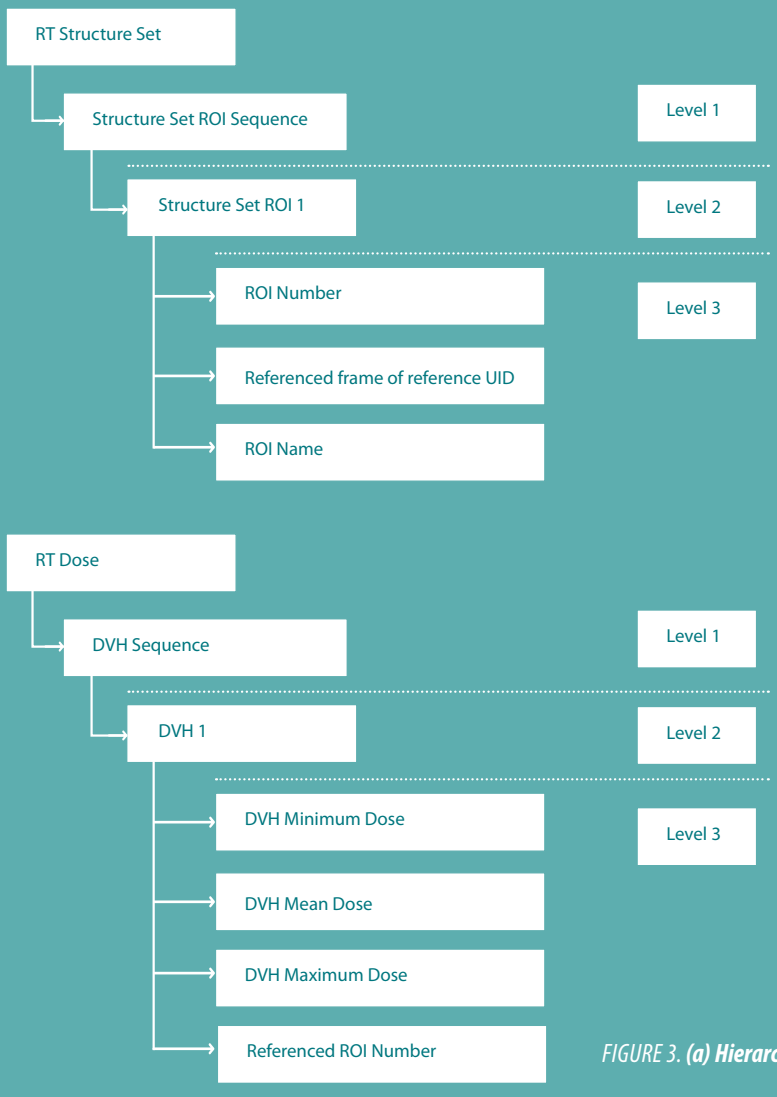


FIGURE 3. ***(a) Hierarchy in RTSTRUCT and (b) RTDOSE***

2.2. Data set

The primary source of data for this study is a proprietary data set composed of anonymized patient information. The dataset was collated for patients who were diagnosed for cancer and underwent radiotherapy, containing CT and DICOM RT files. The corresponding biopsy data of the actual tumor size per patient was also collected in an excel file. We tested our approach in two hypothetical use cases:

Case 1

The first case aims to store and represent the DICOM RT metadata into RDF by flattening the hierarchy in RTSTRUCT and RTDOSE files and retrieve the information via a SPARQL query.

Two patients (including RTSTRUCT and RTDOSE data along with CT data) are used. The hierarchical structure of RTSTRUCT and RTDOSE files were analyzed, and the attribute values were stored in the form of triples (Figure 4).

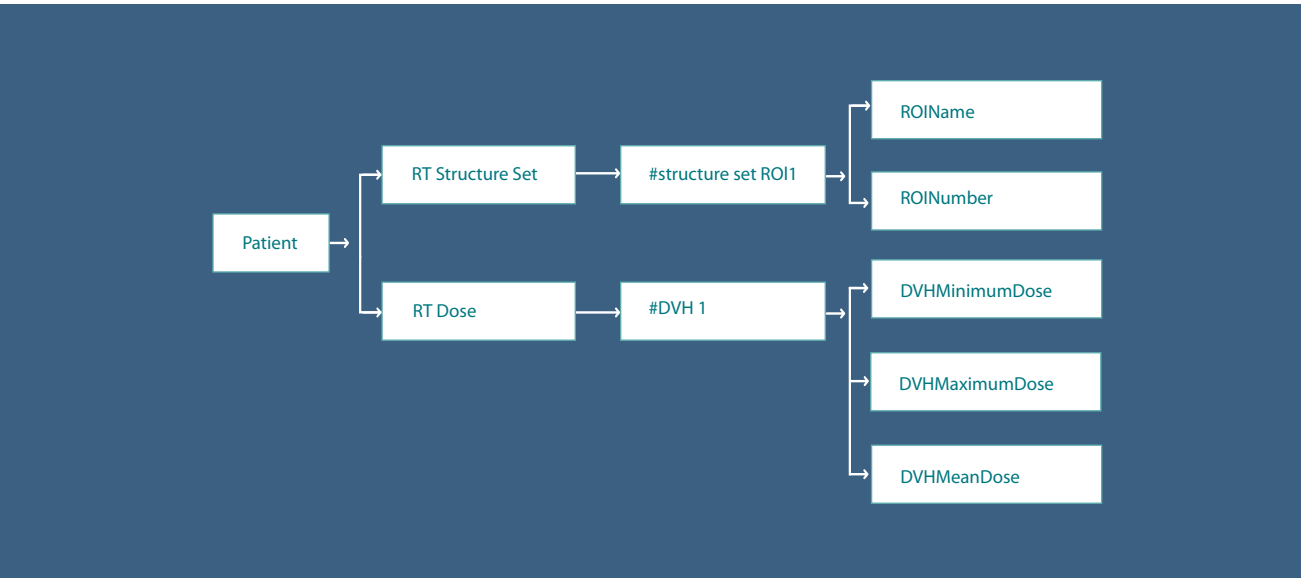


FIGURE 4. ***Graph for the RT STRUCT and RT DOSE***

By creating a common node between RTSTRUCT and RTDOSE to other attributes such as ROI Name, DVH Minimum Dose, the hierarchy is reduced to only one level (Figure 5). The mapping of dose sequence and structure sequence to a reference node makes it easily accessible through SPARQL.A SPARQL endpoint is an API used for querying.

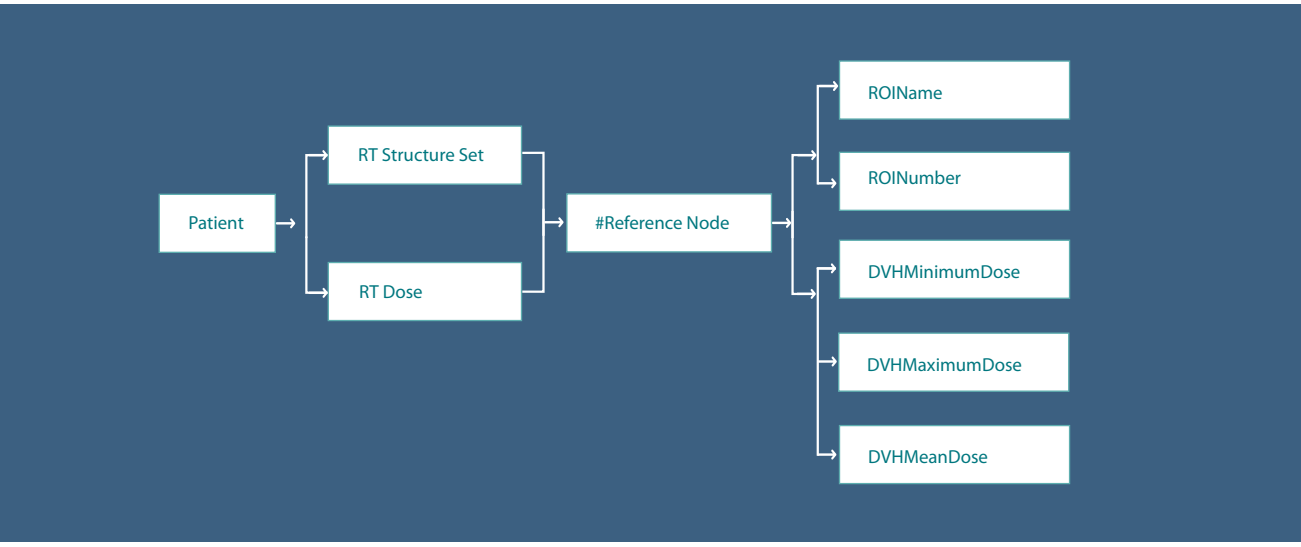


FIGURE 5. ***Graph of RT STRUCT and RT DOSE with a common node***

We further developed a software prototype, where the semantic parser converts the DICOM data into triples and store it into RDF repository. These triples are exposed and easily accessible via a SPARQL endpoint. Such an endpoint enables the users to query a graph by writing a SPARQL query. The prototype architecture is as shown in Figure 6. The open Sesame framework [17] is used for querying and analyzing RDF data. This study has been evaluated qualitatively and quantitatively on 25 varieties of different queries. Evaluation results are shown in the result section and discussed in detail.

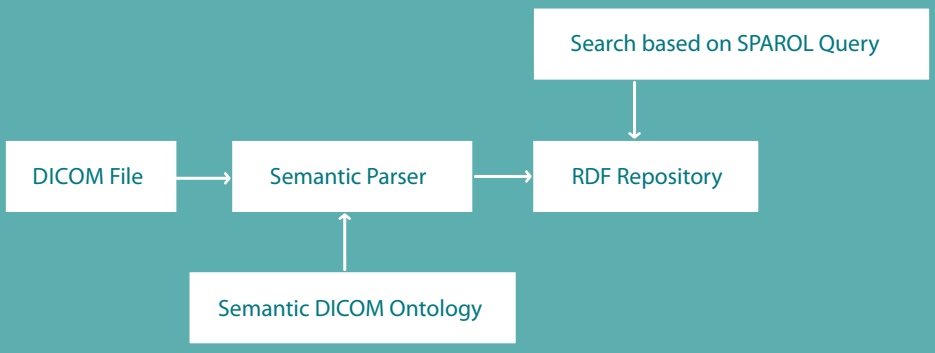


FIGURE 6. ***Architecture of the prototype***

Case 2

The second case aims to store and represent the DICOM RT metadata from RTSTRUCT files into RDF and add the corresponding tumor size information from the biopsy information, based on the fact that literature shows that tumor size is an important factor for making any decision during treatment and is usually stored in pathology reports.

This case was analyzed using 20 patients (RTSTRUCT and tumor size data along with CT data). The RDF store was extended to include the tumor size information as triples. The researchers used the Radiation Oncology Ontology (ROO) to represent the tumor size information and linked it to the patient as an observation procedure finding.

The architecture of the prototype was the same as in case I. This case was evaluated for 10 different queries. Evaluation results are shown in the result section and discussed in detail.

Search Linked RDF Data using a SPARQL Query

Triples enable the users to query a RDF graph by writing a SPARQL query. SPARQL is the RDF query language, able to retrieve and manipulate data stored in RDF format. It allows the user to write the queries against data, by defining "key terms".

We linked the data of RTSTRUCT with RTDOSE objects to identify the DVH dose values and their corresponding ROI Name. Sample queries such as the following were formulated (Figure 7)

- Show the list of all patients
- Get me the list of all patients with DVH Minimum Dose
- Get me the list of patients with DVH Mean Dose and DVH Maximum Dose
- Give the list of all organs
- Get me the list of patients with DVH Mean Dose and their corresponding organ

```
PREFIX Dicom: <http://semantic-dicom.org/dcm#>
SELECT ?organ ?mindose
WHERE {
?x  Dicom:ROIName ? organ.
?x  Dicom:DVHMinimumDose ? mindose
}


SELECT? ID? Structure ?Dose
WHERE{
?x          :Patient ID ?ID;
            [Sequence]  ? Node.
? Node            :ROI Name   ?Structure;
            :Mean Dose        ?Dose. }
```

FIGURE 7. ***Sample SPARQL Queries***

3.2 Natural Language Query Processing for SPARQL Generation

In this section, we explored the use of natural language for querying over RDF data. This required the translation of natural language queries to SPAQRL queries. If a user runs a query like "show me the list of organs for all patients", the parser is required to identify the key terms and relationships and build a SPARQL query.

In the example, organ and patient are the key terms and the building blocks of the SPARQL query.

A simple natural language parser was developed, that works in the context of RT Objects and can understand various queries related to RTSTRUCT and RTDOSE. It works with a customized dictionary comprising limited vocabulary (nouns/entities, phrases, etc.) that caters to the RT Objects context, and used Finite Automata[19] to build the query engine. In order to automate the SPARQL query generation, the query engine already knows (through an ontology)the relationship between doses, organs, patient and some other RT Objects parameters. This relationship was used to automate the query.

For example, when a user enters a text "Show me the mean doses of organs of all *patients", the NLP query engine uses finite automation to establish the correctness of the query. The vocabulary defined in the custom dictionary identifies the nouns and phrases and then converts them into entities in the following sequence of steps:

- "mean dose", "organs", "patients"
- mean dose, Organ, Patient
- DVHMeanDose, ROIName, Patient

The relationship between these three nodes is retrieved and then the SPARQL query is generated out of the relationship. The open Sesame framework workbench and dotNETtoRDF libraries were used to store the RDF files and query the triple data store.

3.0 RESULTS

The flattening of the hierarchal structure of DICOM-RT objects and converting it into RDF triples was successfully implemented and the resulting triples were stored in a RDF repository. The concepts and their relationships were maintained in the RDF graph as they were defined in the SEDI ontology. The proposed method helps to link data of RTSTRUCT and RTDOSE via a reference node, so while executing a query for patients with dose and their corresponding organs, patient lists with values related to those tags were displayed.

When performing a query for patient with DVHs and their corresponding organs, all relevant doses and structures value were returned. These results were manually verified for the correct answers to the questions posed. For case 1, the researchers evaluated 25 different queries where information about doses and structures were retrieved. And for the second case10 different queries were evaluated and the information about structures and patient related tumor size were retrieved.

4.0 DISCUSSION

We have successfully implemented a semantic parser to model the DICOM-RT object tags into RDF triples by reducing the DICOM hierarchy and data mine by executing SPARQL queries, thus confirming our hypothesis. The use of semantics in medical domain is not often explored and the proposed method for semantic search of medical image and information is unique and novel to the best of our knowledge. .

This study focused on a few key attributes for radiotherapy present in DICOM files and pathology reports which are relevant for clinicians. The advantage of the proposed methodology is the ease of information retrieval compared to other techniques. Furthermore, the semantic representation of data in an RDF repository using different ontologies is more flexible than relational databases.

This research work can be extended beyond the modelling of DICOM-RT objects by integrating different ontologies and data sources so that data can be linked enabling seamless information retrieval by computers. Furthermore, the retrieval of data from RDF repositories by writing natural language queries automates the generation of SPARQL queries which may help adoption in future especially by non-experts. . Currently the prototype works for a very limited set of queries and covers only a limited range of linguistic variability in natural language questions. In future work, we aim to improve the engine to allow handling more complex queries.

5.0 CONCLUSION

Efforts in informatics are focused on the structural representation of medical data particularly in oncology, where radiotherapy plays a major role. We performed preliminary tests to evaluate the effectiveness of Semantic Web Technology to store radiotherapy data from DICOM files into RDF graphs and tested the retrieval of data by using natural language to generate SPARQL queries which match the graph pattern.

To show the effectiveness, we implemented a system for semantic modelling of DICOM-RT objects and stored the resulting triples in an RDF repository. Tumor volume information like Gross tumor volume (GTV), Clinical target volume (CTV) from DICOM-RT files were successfully extracted, accessed, and analyzed using Sematic Web Technology. This might enable applications in e.g. the tumor board process, general information mining and longitudinal studies.

The future step would be to develop a system to seamlessly extract and store patient medical record, histopathology and radiotherapy data from the existing data repositories in hospitals and enable easy end-user data mining leveraging Semantic Web Technologies.

REFERENCES

1. *Michael J et.al. The role of radiation therapy in the management of lung, prostate and colorectal cancer in South Dakota. South Dakota journal of medicine 2010; 60-66.*
2. *Nuyts, Sandra. "Use of Imaging Data in Radiotherapy Planning of Head and Neck Cancer: Improved Tumour Characterization, Delineation and Treatment Verification." Head and Neck Cancer Imaging. Springer Berlin Heidelberg, 2006. 345-359.*
3. *The National Electrical Manufacturers Association, Digital Imaging and Communications in Medicine (DICOM), NEMA Publications, PS3.1-PS3.12, 2011.*
4. *Law, Maria YY, and Brent Liu. "DICOM-RT and Its Utilization in Radiation Therapy 1." RadioGraphics 29.3 (2009): 655-667.*
5. *Maria et.al. DICOM-RT and Its Utilization in Radiation Therapy. 2009; 29(3):655-667.*
6. *Maria et.al.DICOM-RT–based Electronic Patient Record Information System for Radiation Therapy 2009; 29(4):961-972.*
7. *Data Mining DICOM RT objects for quality control in radiation oncology. Proc. SPIE 8319, Medical Imaging 2012: Advanced PACS-based Imaging Informatics and Therapeutic Applications, 83190Q, February 23, 2012.*
8. *Brickley D, Guha R.V, McBride B, (2014, Feb. 25), W3C RDF Schema 1.1 [Online]. Available: http://www.w3.org/TR/2014/REC-rdf-schema-20140225/*
9. *Patel-Schneider, P. and Fensel, D. layering the semantic web: Problems and directions. In First International Semantic Web Conference (ISWC2002), Sardinia, Italy, 2002; 16–29.*
10. *Van Soest, Johan, Tim Lustberg, Detlef Grittner, M. Scott Marshall, Lucas Persoon, Bas Nijsten, Peter Feltens, and Andre Dekker. "Towards a semantic PACS: Using Semantic Web technology to represent imaging data." Studies in health technology and informatics 205, pp. 166-170, 2013.*
11. *Möller, Manuel, and Saikat Mukherjee. "Context-Driven Ontological Annotations in DICOM Images-Towards Semantic Pacs." In HEALTHINF, pp. 294-299.*
12. *Brunnbauer, Michael. "DICOM metadata as RDF." In GI-Jahrestagung, pp. 1796-1804. 2013.*
13. *Sauermann, L., Cyganiak, R., & Völkel. Cool URIs for the semantic web. 2011.*
14. *http://www.w3.org/TR/sparql11-overview/*

15. *Semantic flooding: Search over semantic links. Dept. of Inf. Eng. & Comput. Sci., Univ. of Trento, Trento, Italy, IEEE 26th International Conference on Conference: Data Engineering Workshops (ICDEW), 2010.*
16. *McGuinness, D. L. and van Harmelen, F. (2004). OWL Web Ontology Language overview. W3C recommen- dation,WorldWideWeb Consortium.*
17. *Sesame, Available:http://rdf4j.org/*
18. *Thalheim, Bernhard. (1993). Foundations of Entity-Relationship Modeling. Annals of Mathematics and Artificial Intelligence. 7. 197-256. 10.1007/BF01556354.*
19. *Nagy B.andOtto F. "Finite-state Acceptors with Translucent Letters." DOI: 10.5220/0003272500030013.InProceedings*

CHAPTER 6

AN APPROACH TOWARD AUTOMATIC CLASSIFICATION OF TUMOR HISTOPATHOLOGY OF NON–SMALL CELL LUNG CANCER BASED ON RADIOMIC FEATURES

Ravindra Patil
Geetha Mahadevaiah
Andre Dekker

Tomography. 2016;2(4):374–377. doi:10.18383/j.tom.2016.00244

ABSTRACT

Non-small cell lung cancer (NSCLC) contributes to 85% of all lung cancer burden. The histology of the tumor (squamous cell carcinoma, large cell carcinoma, and adenocarcinoma and "not otherwise specified") has prognostic significance and it is therefore imperative to identify tumor histology for personalized medicine, but biopsies are not always possible and carry significant risk of complications. In this study we have used Radiomics, which provides an exhaustive number of informative features, to aid in diagnosis and therapeutic outcome of tumor characteristics in a noninvasive manner. This study evaluates radiomic features of NSCLC to identify the histopathology of the tumor. We have included 317 subjects and classified the underlying tumor histopathology into its four main sub types. The performance of the current approach is 20% more accurate than an approach just considering volumetric and shape based features.

Keywords

LUNG CANCER
ORGANS AT RISK
RADIOTHERAPY
RADIOMICS
NON-SMALL CELL LUNG CANCER (NSCLC)

1.0 INTRODUCTION

Non-small cell lung cancer accounts for 85% of all the lung cancers and it is the second most common cause of cancer in both men and women. Estimates by the American Cancer Society for 2016 are an incidence in the United States of 224,390 new cases of lung cancer (117,920 in men and 106,470 in women) and 158,080 deaths from lung cancer (85,920 in men and 72,160 in women). Every year, more people die of lung cancer than of colon, breast and prostate cancer combined together. 2 out of 3 people diagnosed with lung cancer are 65 or older, while less than 2% are younger than 45 years [(1)]. The different factors that can affect survival include genetic factors, clinical factors such as age, and overall health, the size and stage of the tumor and the histological subtype of NSCLC. With regards to the subtype, many studies have identified a link between the subtype and survival. For example, Ma et al [(2)] recently showed that adenocarcinoma patients have a worse prognosis. Similarly, Yano et al [(3)] showed recently squamous cell carcinoma have worse outcomes and concluded that the surgical management should be different for different subtypes. Given this knowledge, it is imperative to have histological classification of non-small cell-lung carcinoma. Currently, a biopsy is performed to identify the subtype which involves an invasive procedure and frequent sample extraction from the site of interest. Also, it is painful, costly and not without risk [(4)].

In recent years more emphasis has been given to non-invasive diagnosis and screening of lung cancer [(5, 6)]. An advanced technique named "Radiomics" involving the extraction of large quantitative features, which results in conversion of images into higher dimensional minable data, can be used for building decision support systems. This is in contrast to the traditional practice of treating medical images as pictures used solely for visual interpretation. The output from a Radiomics analysis contains first, second, and higher-order statistics data derived from the entire image or a particular Region of Interest (ROI).It can be performed with tomographic images from CT, MR imaging, and PET studies [14].The seminal work performed by Aerts et.al [(7)] showed radiomic features having prognostic power in lung and head-and-neck cancer patients. Also, there exist several studies in NSCLC by Al-Kadi et.al, Cook et.al and Ganeshan et.al [(8-10)] in which texture analysis is applied to predict lung cancer outcomes based on factors associated with tumor texture. In this work we hypothesize that radiomic texture features can classify the different subtypes of NSCLC in a noninvasive approach.

The aim of this study is to perform a radiomic analysis to identify tumor histopathology in NSCLC and classify into squamous cell carcinoma, adenocarcinoma, large cell carcinoma and not otherwise specified (NOS) non-small cell lung cancer. We extracted radiomic features and evaluated their ability to classify the tumor histopathology and also evaluated how these features compare with non-radiomic, "normal" features (e.g., volume, diameter)

2.0 METHODOLOGY

The applied methodology involves image acquisition, segmentation, feature extraction and model validation to identify the histology of the under lying tumor. Figure 1 summarizes the above mentioned approach. The approach involved was to characterize the tumor region through extraction of radiomics features which earlier were thought to be redundant or non-useful for clinical outcome. The obtained CT image is segmented to define the tumor region and in this study Gross Tumor volume (GTV) is used. The segmented region is used to extract features based on tumor intensity, shape and texture. The extracted features are analyzed to build a decision support system to identify histopathology of the tumor.

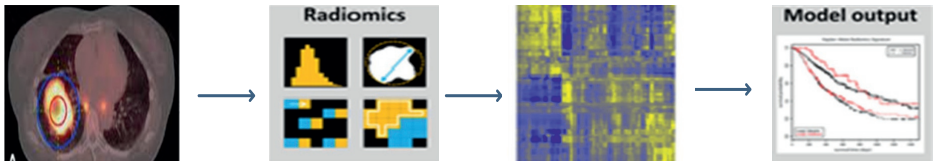


FIGURE 1. ***The Radiomics workflow, images partly from (7) with permission***

Data

The data used in this study were obtained from The Cancer Imaging Archive (TCIA) (http://www.cancerimagingarchive.net/) from the collection of NSCLC-Radiomics. In all 317 patient CT data having NSCLC cancer were considered out of which 40 were Adenocarcinoma, Large cell carcinoma – 108, Squamous cell carcinoma – 110 and "not otherwise specified" were 59. The demographic distribution is mentioned in Table 1. The ground truth of the tumor region was provided by the radiation oncologist in the DICOM-RTSTRUCT file and hence no explicit segmentation algorithm was used for the tumor regions identification. The data was not preprocessed or normalized so as not to devalue any of the clinical information present in the images.

Subject Characteristics	Adenocarcinoma	Large cell carcinoma	Squamous cell carcinoma	NOS
Number of subjects	40	108	110	59
Male	20	65	70	41
Female	20	43	40	18
Mean Age (years)	67.2	66.9	70.2	65.6

TABLE 1. ***Distribution of the subject characteristics***

Feature extraction

An important sub step of Radiomics is the high throughput extraction of quantitative image features. In total 431 imaging features were extracted from the site of the tumor identified by the manual delineation performed by the radiation oncologist. The feature extraction algorithm was custom implemented in Matlab. The quantitative imaging features were divided into four sub groups namely (i) First order statistics (ii) Shape and size based features (III) Textural features (IV) Multiscale wavelet; which describes tumor phenotypes. First-order statistics provides distribution of voxel intensities within the CT image through commonly used and basic metrics (e.g. energy, entropy, kurtosis).

The shape and size based features provide tumor compactness, volume, area metrics along with how spherical, rounded or elongated the tumor is. The first order statistics represents only the information related to the gray level distribution and doesn't provide pertinent information on the relative position of various gray level across images. Then, we computed textural features using gray level co-occurrence (GLCM) and gray level run-length (GLRLM) texture matrices. All the voxel intensities were resampled into equally spaced bins using a width of 25 Hounsfield Units. The former step reduces image noise as well as normalizes the intensities and will allow direct comparison of all calculated textural between patients. The matrices were determined considering 26 - connected voxels in all 13 directions in 3D.

The wavelets features help to decouple original images into high and low frequencies. In this analysis, we applied one level un-decimated 3-D wavelet transform on the GTV. The original image is decomposed into 8 level of wavelet decomposition (X_{LLL}, X_{LLH}, X_{LHL}, X_{LHH}, X_{HLL}, X_{HLH}, X_{HHL} and X_{HHH}), where L and H are low pass and high pass. For example, X_{HLH} is interpreted as the high-pass sub band resulting from directional filtering in x with high pass, a low pass along y- direction and high pass in z-direction.

$$X_{LLH}(i,j,k) = \sum_{p=1}^{N_H} \sum_{q=1}^{N_L} \sum_{r=1}^{N_H} H(p)L(q)H(r)X(i+p, j+q, k+r)$$

Where $N_H$the length of filter H and N_L is the length of filter L.

For each of the decompositions obtained above we computed first order statistics and textural features as mentioned earlier. The feature vector length for each of the subjects is 431.

The first order statistics accounts to 14 features, shape and size based features were 8, textural features accounted to 33 and remaining 376 features were derived from wavelet decomposition.

Classification

As the classes of data were skewed and in order to eliminate the bias in learning due to unequal samples in each class, the SMOTE algorithm [11] was applied on the dataset. Two data models were built in this experiment, the first being called radiomic model contained all the radiomic features extracted in the above step and the second being the normal model consisting of typical normal features which were subset of radiomic features consisting of (Energy, Entropy, Kurtosis, Maximum, Mean, Mean Absolute Deviation, Median, Minimum, Range, RMSvalue, Skewness, Standard Deviation, Uniformity, Variance, Compactness, Maximum 3D diameter, Spherical Disproportion, Sphericity, Surface Area, Surface to Volume Ratio, Volume). Following which, a multi-category support vector machine using R with package (e1071) [13] was used for classifying the data into predefined classes.

The results obtained were based on 10 fold cross validation on the dataset. The experimental parameters set for SVM were kernel type: rbf, gamma: 1.0, C: 0.1, eplison: 10^(-4).The selection of the above mentioned hyper parameters were performed based on a grid search to yield the best accuracy value. Further, the ranking of features based on the built model was performed to identify the most important predictive features that aided in the classification using the Caret package in R.

3.0 RESULTS

The demographic of the dataset consist of both male (60%) and female (40%) subjects. The average age of the pool was 68 years. The metrics of the histology classification obtained by considering radiomics features and that of normal features is shown in the Figure 2. It can be observed that there is 20 percent improvement ($p < 1.2e-4$) over the accuracy in identification of the tumor histopathology using radiomic features compared to normal features.

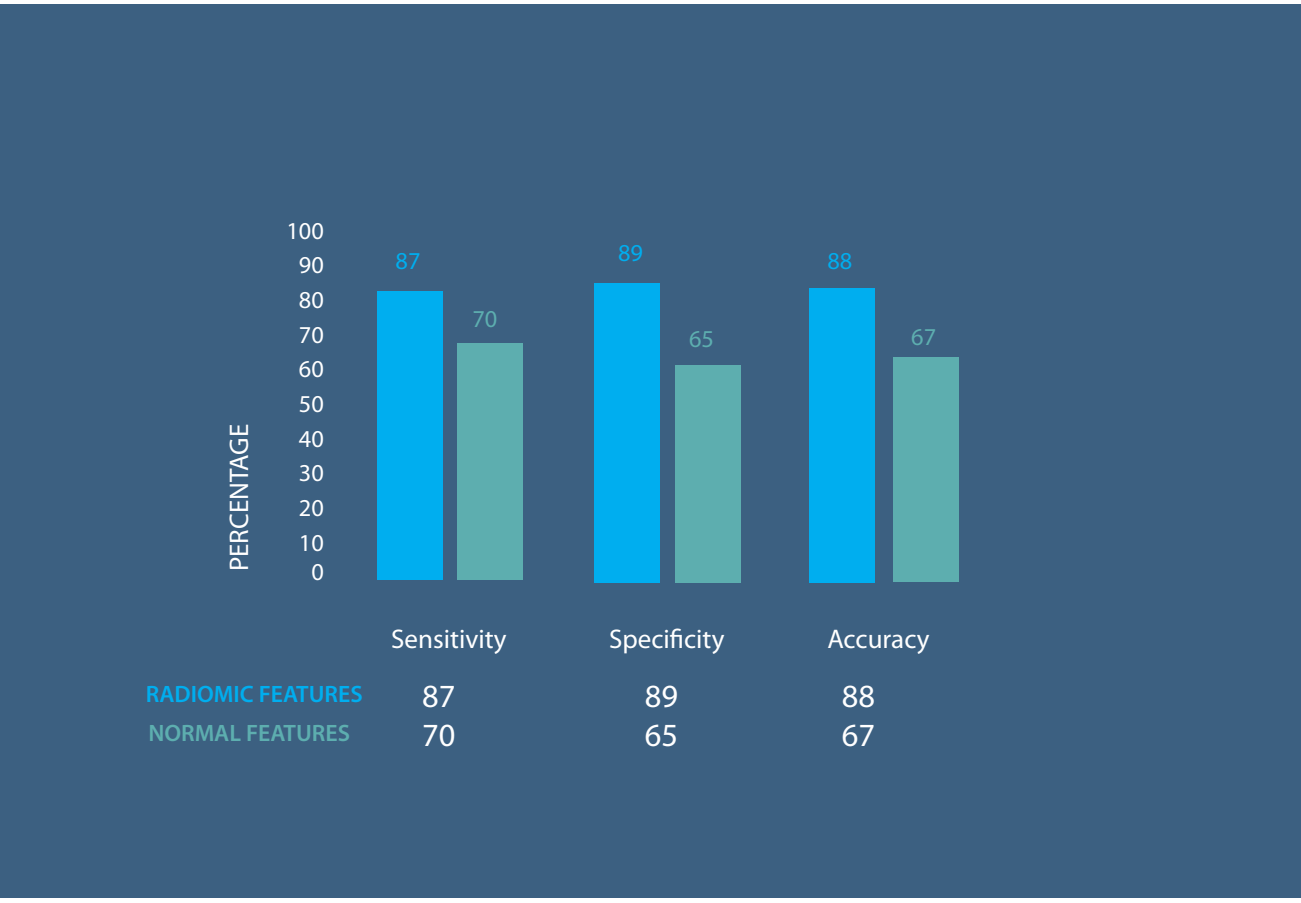


FIGURE 2. ***Classification metrics using radiomic and normal features***

Also, feature ranking was employed on the radiomic model to rank the features based on the importance value. The top 40 radiomic features which aids in classification of the tumor histopathology are shown in Figure 3. It is interesting to note that volume as an independent feature doesn't rank as the top contributor for the classification. Also, wavelet based features dominate top contributing features for the classification, strengthening the claim that additional features aid in better information extraction. Hence, radiomics is able to quantify phenotypical differences from medical images by using a large set of imaging features and providing more hidden information compared to normal features in classification of tumor histopathology.

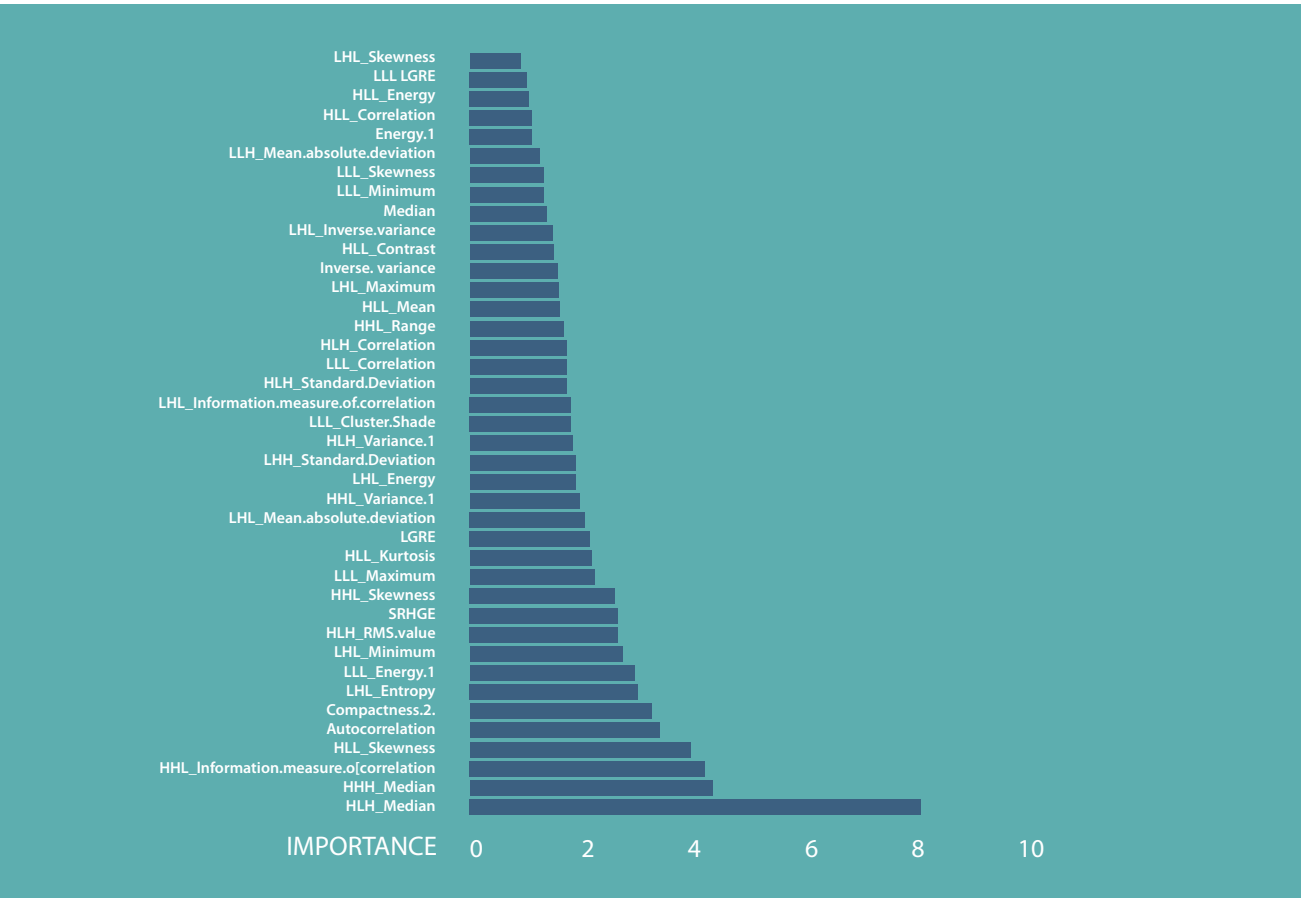


FIGURE 3. ***Radiomic features ranking based on importance***

4.0 DISCUSSION

In this study we have adapted radiomics to classify NSCLC tumors namely adenocarcinoma, squamous cell carcinoma, large cell carcinoma and not-otherwise-specified by extracting imaging features from annotated tumor region. The results of study show that the radiomics has greater potential than normal features in classifying the tumor histopathology. Our approach provides a noninvasive, fast, low cost and repeatable way of investigating tumor histology, hence speeding up the development of personalized medicine. However, our method directly considers the tumor region segmented by the radiologist, in future our work will be focused on integrating the segmentation in the workflow. Further the limitations of this study are a lack of validation by an external dataset and that we did not compare it to non-invasive liquid biopsies [12], which might also provide a method to determine histology although this is still quite an experimental technique.

REFERENCES

1. *A report on Lung Cancer (Non-Small Cell), American Cancer Society, 2016 (http://www.cancer.org/acs/groups/cid/documents/webcontent/003115-pdf.pdf).*
2. *Ma LH, Li G, Zhang HW, Wang ZY, Dang J. The effect of non-small cell lung cancer histology on survival as measured by the graded prognostic assessment in patients with brain metastases treated by hypofractionated stereotactic radiotherapy. Radiat Oncol. 2016 Jul 13;11:92. doi: 10.1186/s13014-016-0667-x.*
3. *Yano M, Yoshida J, Koike T, Kameyama K, Shimamoto A, Nishio W, Yoshimoto K, Utsumi T, Shiina T, Watanabe A, Yamato Y, Watanabe T, Takahashi Y, Sonobe M, Kuroda H, Oda M, Inoue M, Tanahashi M, Adachi H, Saito M, Hayashi M, Otsuka H, Mizobuchi T, Moriya Y, Takahashi M, Nishikawa S Matsumura Y, Moriyama S, Fujii Y. The Outcomes of a Limited Resection for Non-Small Cell Lung Cancer Based on Differences in Pathology. World J Surg. 2016 Jun 30. doi:10.1007/s00268-016-3596-9.*
4. *Wu CC1, Maher MM, Shepard JA. Complications of CT-guided percutaneous needle biopsy of the chest: prevention and management. AJR Am J Roentgenol. 2011 Jun;196(6):W678-82. doi: 10.2214/AJR.10.4659.*
5. *Bajtarevic A1, Ager C, Pienz M, Klieber M, Schwarz K, Ligor M, Ligor T, Filipiak W, Denz H, Fiegl M, Hilbe W, Weiss W, Lukas P, Jamnig H, Hackl M, Haidenberger A, Buszewski B, Miekisch W, Schubert J, Amann A. Noninvasive detection of lung cancer by analysis of exhaled breath. BMC Cancer. 2009 Sep 29;9:348. doi: 10.1186/1471-2407-9-348.*
6. *D'Urso V, Doneddu V, Marchesi I, Collodoro A, Pirina P, Giordano A, Bagella L. Sputum analysis: non-invasive early lung cancer detection. J Cell Physiol. 2013 May;228(5):945-51. doi: 10.1002/jcp.24263.*
7. *Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies & Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications 5, Article number: 4006 (2014) doi:10.1038/ncomms5006.*
8. *Al-Kadi OS, Watson D. Texture analysis of aggressive and nonaggressive lung tumor CE CT images. IEEE Trans Biomed Eng. 2008; 55(7):1822–1830. doi: 10.1109/TBME.2008.919735.*

9. *Gary J.R. Cook, Connie Yip, Muhammad Siddique, Vicky Goh, Sugama Chicklore, Arunabha Roy, Paul Marsden, Shahreen Ahmad,David Landau. Are Pretreatment 18F-FDG PET Tumor Textural Features in Non–Small Cell Lung Cancer Associated with Response and Survival After Chemoradiotherapy? J Nucl Med 2013; 54:1–8 doi: 10.2967/jnumed.112.107375.*

10. *Ganeshan B, Goh V, Mandeville HC, Ng QS, Hoskin PJ, Miles KA. Non-small cell lung cancer: histopathologic correlates for texture parameters at CT. Radiology. 2013 Jan;266(1):326-36. doi: 10.1148/radiol.12112428. Epub 2012 Nov 20.*

11. *Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique,SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16 (2002) 321–357.*

12. *Graham Brock, Elena Castellanos-Rizaldos, Lan Hu, Christine Coticchia, Johan Skog. Liquid biopsy for cancer screening, patient stratification and monitoring. Transl Cancer Res 2015;4(3):280-290 .*

13. *https://cran.r-project.org/web/packages/e1071/e1071.pdf*

14. *Jun Wang, Xia Liu, Di Dong, Jiangdian Song, Min Xu, Yali Zang, Jie Tian. Prediction of malignant and benign lung tumor using a quantitative radiomic method. 38th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC), 2016.*

CHAPTER 7

FRACTAL ANALYSIS IN HISTOLOGY CLASSIFICATION OF NON-SMALL CELL LUNG CANCER

Ravindra Patil
Geetha Mahadevaiah
Srinidhi Bhat
Dinesh M.S
Leonard Wee
Andre Dekker

Medical Imaging: Artificial Intelligence, Image Recognition, and Machine Learning. Chapter 4: Page 63

ABSTRACT

Non-small cell lung cancer (NSCLC) accounts for 85% of all the lung cancers. Non-invasive identification of the histology of NSCLC aids in determining the appropriate treatment approaches. In this study we have studied the usage of Radiomics with application of fractals to arrive at the histology classification of NSCLC using Lung CT images. This study suggests that fractals can play a vital role in Radiomics, providing information not only about the Gross Tumor Volume (GTV) structure, but also can help in characterization of the tumor. It was observed that Fractal dimension based features were among the top 15 features that aided in histology classification based on 317 subjects and improved the existing classification accuracy of the NSCLC histology by 8% by adding fractal dimension as a feature.

1.0 INTRODUCTION

In both sexes combined, lung cancer is the most commonly diagnosed cancer(11.6% of the total cases) and the leading cause of cancer death. The total number of lung cancer cases in 2018 alone amounted to 2,093,876 and number of deaths with lung cancer being 1,761,007. Non-small cell lung cancer (NSCLC) accounts for 85% of all the lung cancers [1]. The cause and the survival of NSCLC subject varies across age, genetic profile, size of tumor and histopathology of tumor. There are various studies that have established a correlation between the subtypes of NSCLC (squamous cell carcinoma, large cell carcinoma, and adenocarcinoma and "not otherwise specified") to that of the survival. Also, it was observed that the adenocarcinoma prognosis is poor compared to those of non-adenocarcinoma [2]. It was also concluded that the surgical management should be different for each sub categories of NSCLC [3]. The current approach of sub type detection is performed using a biopsy procedure, where the tissue under observation is biopsied to determine the subtype, which is invasive in nature. The invasive approach is painful, costly and also it's not devoid of complications [4]. Recently, several studies have been undertaken to identify the sub categories of NSCLC non-invasively using an approach of Radiomics, wherein a large amount of quantitative features are mined and decision support models are built to achieve the desired objective [5]. Of late, Radiomics has been applied to several medical problems such as tumors of lung, breast and prostate. Radiomics has also been applied on images extracted from different medical imaging techniques (computed tomography (CT), magnetic resonance (MR), and positron emission tomography (PET) [6-9, 16-19], showing promising results in each case.

Also, there has been lot of interest in application of Fractals in the oncology domain. Fractals are mathematical objects which have a non-integer dimension. These object manifests repeating pattern at different size scales; this property is quantified by a parameter named fractal dimension that measures the self-similarity grade of the structure under analysis [10]. Such mathematical objects can be self-similar and resembles the repeated pattern within itself. These patterns have been studied in the oncology domain to differentiate between malignant and benign tumors in case of breast cancer. There havebeen comprehensive reviews on the use of fractal dimensions in various medical research areas such as pathology [11 -12]. Recent trends show fractals

to be a useful measure of the pathologies of the vascular architecture, tumor/parenchymal border, and cellular/nuclear morphology. A procedure that combines fractal and segmentation analysis has been proposed to investigate heterogeneity in cancer cells on MR images, similar to the approach described by Szigeti et al.,[20] for studying lung tumor heterogeneity on mice CT scans. More details on the fractals and its applications in oncology can be found in [11].

Other major studies include different fractal measures such as the power-law behavior of the Fourier spectrum of gray-scale images which was used by Heymans et al. [13] to characterize the microvasculature in cutaneous melanoma. The fractal dimension quantified the degree of randomness in the vascular distribution, a characteristic that cannot be easily captured by the vascular density. Another study by Cusumano et al.[14] [21]aimed to use a fractal based radiomic approach to predict complete pathological response after chemo-therapy in locally advanced rectal cancer (LARC). Fractals played an important role giving information not only about the gross tumor volume (GTV) structure but also describing information about the sub-populations in the GTV [14].

In this study, we investigate the role of Fractals in classification of Non-Small Lung Cancer histology. We follow the box counting approach that aims to overlay boxes of various sizes over the area of interest and to optimize this quantity with the size of each box. Overall, our technical contributions are:

1) Building an algorithm which can compute the Fractal Dimension of a two Dimensional Region of Interest in a given volume.

2) Analyzing the extent to which fractal dimension aids in classification of NSCLC subtypes in the presence of other radiomic features.

2.0 METHODOLOGY

The methodology adapted in this study consists of acquisition of the NSCLC images with varied histology subtypes (i.e. squamous cell carcinoma, large cell carcinoma, and adenocarcinoma and "not otherwise specified"), segmentation of the GTV, computation of fractals, extraction of radiomic features and validation of the model to identify the histology of the underlying tumor. The pictorial depiction of the workflow is shown in the Figure 1.

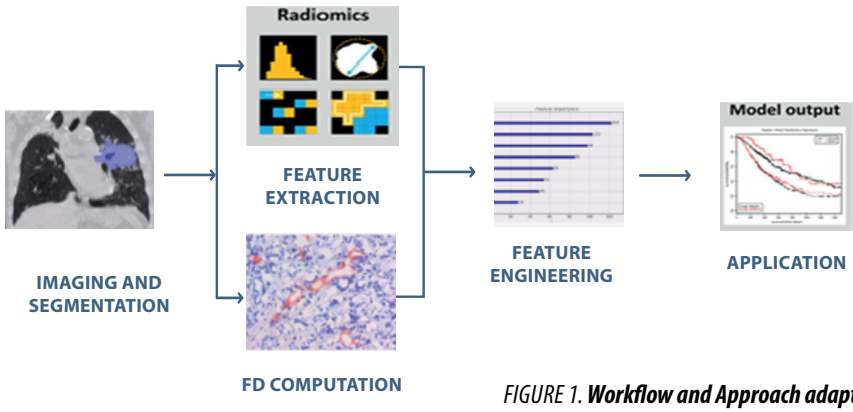


FIGURE 1. ***Workflow and Approach adapted***

Image Analysis

The following experiment involved Computerized Tomography (CT) Images of NSCLC. Images of 317 patients were used in this study and the data set was obtained from the (http://www.cancerimagingarchive.net/) [9] from the collection of NSCLC-Radiomics. The images were in Digital Imaging and Communications in Medicine (DICOM) format and the CT image of each patient had the corresponding Structural Report file (RTSTRUCT) which consist of GTV delineation performed by a team of expert radiologists. The distribution of the demographic of data is depicted in the Table 1.

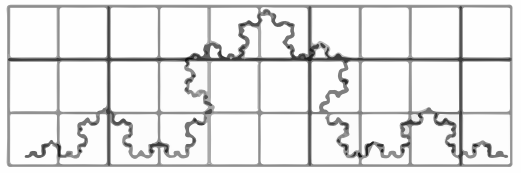
SUBJECT CHARACTERISTICS	ADENOCARCINOMA	LARGE CELL CARCINOMA	SQUAMOUS CELL CARCINOMA	NOS
Number of subjects	40	108	110	59
Male	20	65	70	41
Female	20	43	40	18
Mean Age (years)	67.2	66.9	70.2	65.6

TABLE 1. ***Subject Demographics***

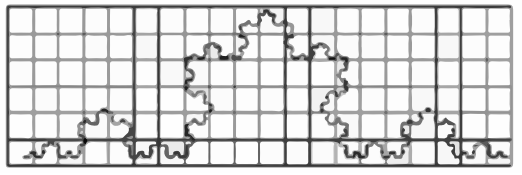
Accessing the DICOM tags which contained the contour information of the Gross Tumor Volume (GTV) using the information in the RTSTRUCT file, we create a mask depicting the GTV. The extracted mask was superimposed on the actual image to delineate the region of interest. Further, the minmax normalization of the images was performed to ensure the effect due to spike pixels are minimized.

Computation of Fractal Dimension

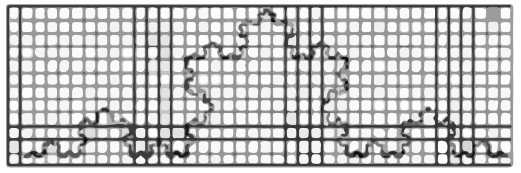
The fractal dimension was computed based the box counting approach and was further optimized to deduce the fractal dimension for each region of interest. In our approach we extracted every surface contour of the GTV and box counting algorithm was applied to compute fractal dimension on each of the slice containing the tumor volume. The pictorial representation of the approach can be seen in the Figure 2 with varied grid sizes [15]



The Koch curve with unit 1 grid size, with 18 containing the curve



The Koch curve with unit 1/2 grid size, with 41 containing the curve



The Koch curve with unit 1/4 grid size, with 105 containing the curve

FIGURE 2. ***Sample representation of box counting approach with varied N [15]***

In this study the FD was computed on the impinged GTV on the CT slices. The equation for the computation is mentioned in (1).

$$FD = \frac{\log(N)}{\log(\frac{1}{r})} \quad (1)$$

Where:

FD is the Fractal Dimension
N is the number of boxes needed to cover the region of interest
r is the size of each box

The above equation is recursively applied by varying the size of the box, there by converting it into a curve fitting solution. In principle, this corresponds to a line fitting problem, where points corresponding to the N and 1/r are fit and the slope of line provides the fractal dimension. This is one of the reasons we have applied a logarithm to equation [1]. This reduces a curve fitting method to a line fitting method, which is computationally simpler to solve. Also, logarithm, being a monotonic function in nature doesn't alter the behavior of the original equation. The FD value for sample subsets are shown in the Figure. 3, which is computed based on each of the slices.

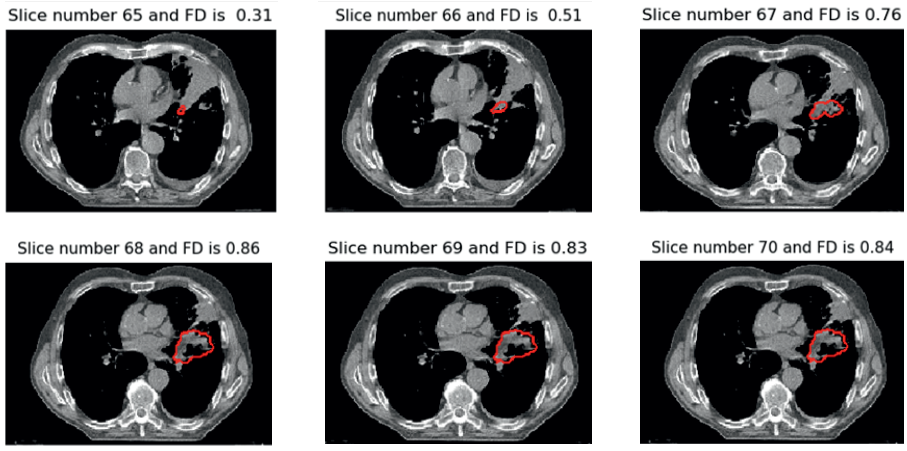


FIGURE 3. ***Subject Lung001 from the NSLC data set with FD computed value***

Extraction of Radiomics Features

The quantitative image features were extracted from the GTV, wherein these imaging features where divided into four sub categories [1] First order statistics [2] Textural Features [3] Shape and Size based features [4] wavelet features. The first order features provides the voxel intensity distribution within GTV. The textural features were computed using gray-level co-occurrence and gray-level run-length texture matrices, which aid in providing the relative position of various gray level distribution. Shape and size features provide information on how spherical, elongated or rounded the tumor manifests and provide information about area , tumor compactness and volume. The wavelet features provide the information by decoupling the region into high and low frequencies with GTV as input. In this approach, the original images are decomposed into 8 level of wavelet decomposition (X_{LLL}, X_{LLH}, X_{LHL}, X_{LHH}, X_{HLL}, X_{HLH}, X_{HHL} and X_{HHH}), where L and H are low pass and high pass. For example, X_{LHL} is interpreted as the low-pass sub band resulting from directional filtering in x with low pass, a high pass along y- direction and low pass in z-direction.

$$X_{LLH}(i,j,k) = \sum_{p=1}^{N_H} \sum_{q=1}^{N_L} \sum_{r=1}^{N_H} H(p)L(q)H(r)X(i+p, j+q, k+r) \quad (2)$$

Where $N_H$the length of filter H and N_L is the length of filter L.

In total 431 Radiomics features were extracted and these formed the feature vector for each of the subject. Further, to this feature vector, Max FD, average FD obtained from the fractal dimension computation were augmented making it a 433 feature vector for each subject

Classification

Two data models were built in this experiment, one with all radiomic features including fractal features extracted and the second with only radiomic features excluding the fractal features. A Random Forest Classifier (RFC) was used to model the multiclass classification problem of predicting the histology of the tumor into one of the following sub categories: Squamous cell carcinoma, Large cell carcinoma, Adenocarcinoma and not otherwise specified. The RFC was implemented using Sklearn ensemble package in python and to obtain the best set of parameters we defined a grid of hyper-parameter ranges using Scikit-Learn's Randomized Search CV package by performing 10-fold cross validation with each combination of values.

The hyper parameters and the values of the RFC tuned in this experiment were:

• **max_features**: This parameter describes the maximum number of features the random forest is allowed to try in individual tree. The value in this experiment was chosen as "auto" which will take all the features in every tree. Though this option decreases the speed of the algorithm, it provides a high number of options to be considered at each node. Moreover, in this experiment we have first chosen the top 15 contributing features and then used them for the classification task, hence the sample space is reduced compared to the original 433.

• **n_estimators**: This parameter describes the number of trees to be built before taking the maximum voting of the predictions. Empirically, in this experiment, the optimal value for this parameter has been 10.

• **Criterion**: This parameter describes the criteria of split. In this experiment we have used the Gini index impurity measure.

• **max_depth**: This parameter represents the depth of each tree in the forest. The deeper the tree, the more splits it has and can capture more information regarding the data. In this case the optimal depth value was found to be 15.

• **min_samples_leaf**: This describes the minimum samples required to be at the leaf node. In this experiment the optimal value of parameter is found to be 3.

3.0 RESULTS

There were a total of 317 subjects that were considered for the histology classification, out which 40% of the subjects were female and the average age was 68 years. Pearson's correlation analysis was performed to understand the relationship of first order Radiomics and FD features with that of the histology class. It can be observed that max FD has maximum correlation with respect to histology class compared to other first order features. Further, it can also be seen that tumor volume as an independent feature ranks much lower than the FD in Figure 4. This is in line with the understanding that the morphology of the tumor provides a vital distinction other than tumor volume in histology differentiation.

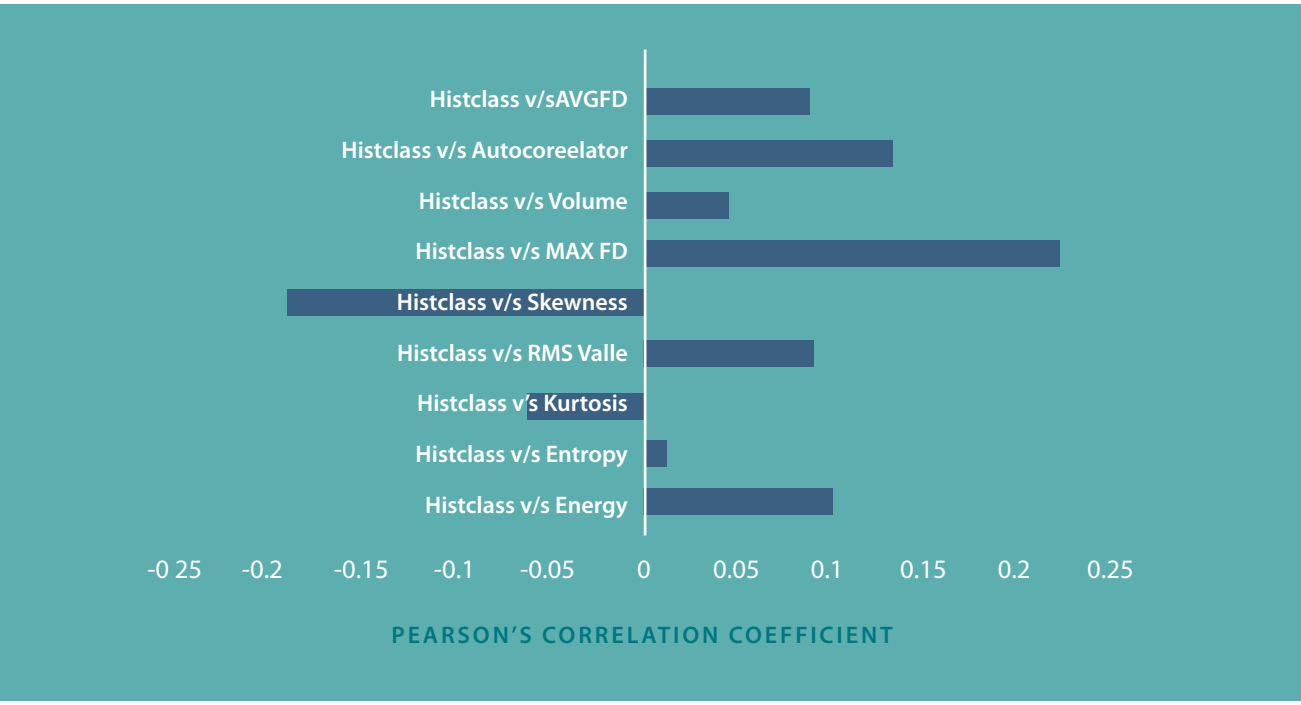


FIGURE 4. ***Correlation of Histology with various first order Radiomics features and FD features***

The classification accuracy considering the Radiomics features accounted for 74% however with inclusion of FD features it was improved to 86 % ($P < 0.001$). Also, there is an improvement of 8 % and 17% in terms for sensitivity and specificity respectively by considering the FD features (Figure 5).

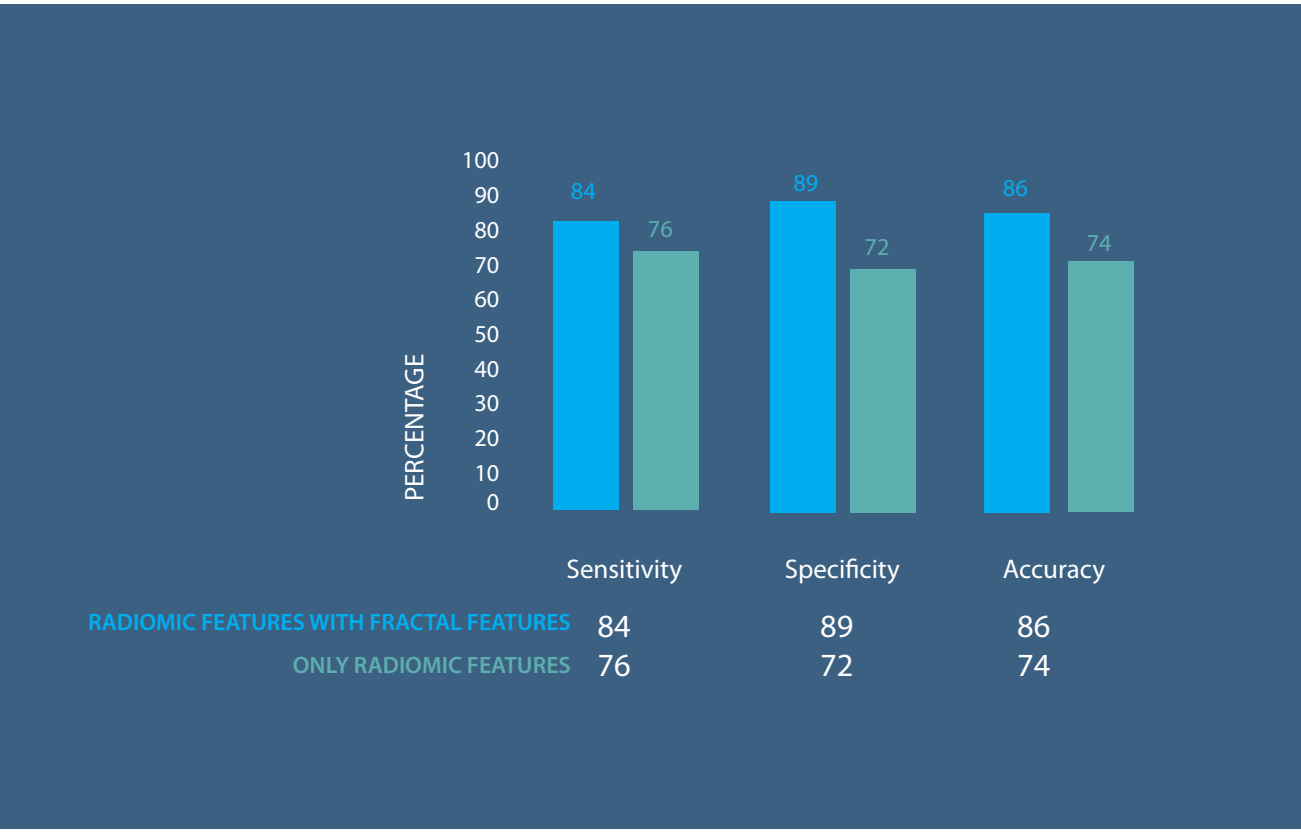


FIGURE 5. ***Classification metric comparison between with FD and without FD features inclusion***

Further, the features were ranked based on a random forest classifier on the priority of importance for classification of histology as shown in Figure 6. The top features contributing toward histological classification using radiomic features include HHH_ Sum Entropy, HHH_LRGE, HHH_RLN, Inverse Difference Moment Normalized (IDMN), HHH_GLN, HHH_SRHGE, HHH_IDMN, Informational measure of correlation 2 (IMC_2), Maximum FD, Average FD, HHH_Sum Variance. Also, it is interesting to note that the FD derived features, Max FD and Avg FD features were ranked among the top 15 features that aids in histology classification of NSCLC.In essence, wavelet-based features and FD parameters dominate as top contributing features for the histology classification.

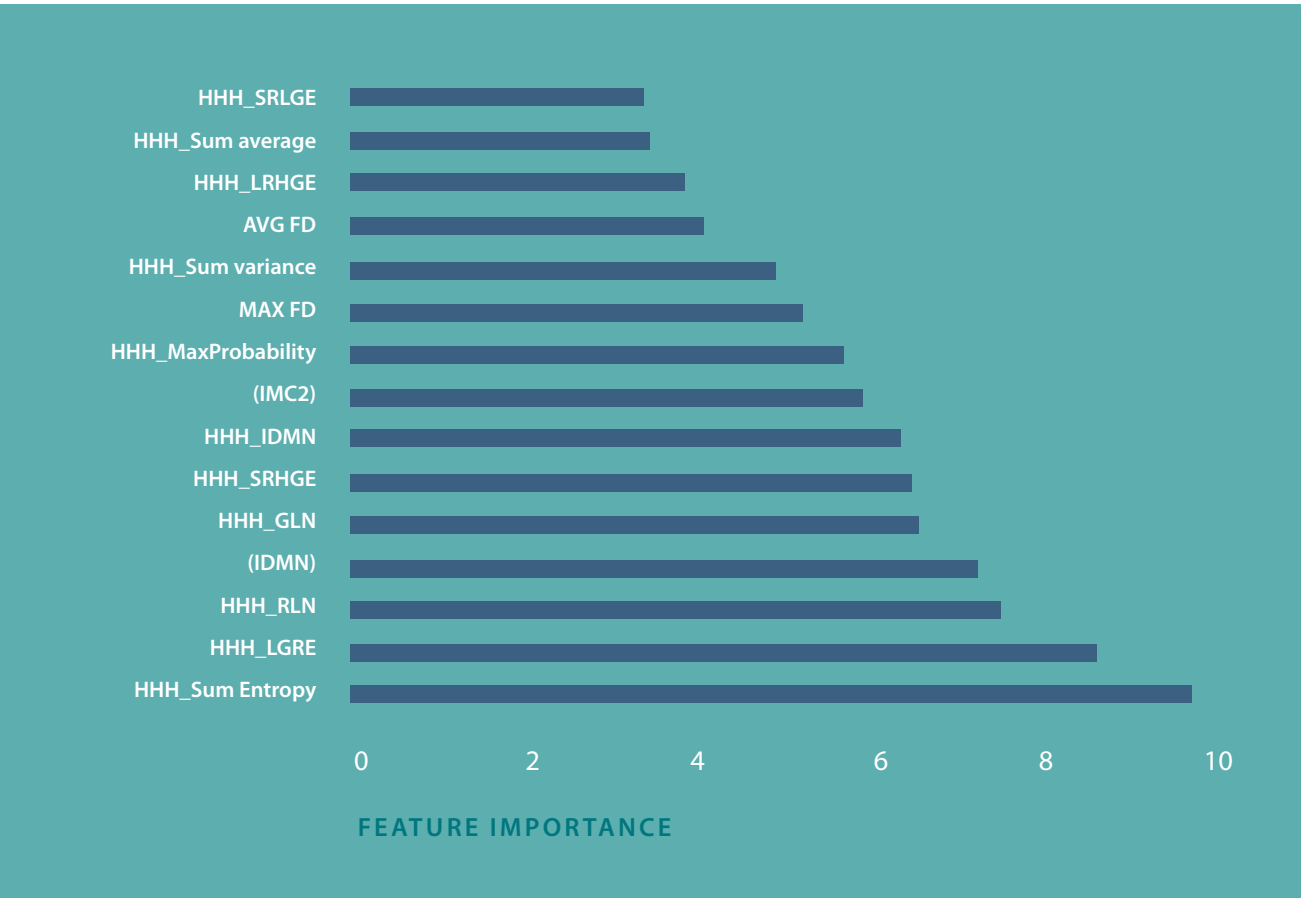


FIGURE 6. ***Feature ranking based on the importance***

4.0 CONCLUSION

In this study, we could establish that the Fractal derived features play an important role in histology classification of NSCLC. Applying fractal analysis on a 2-D contour region can provide valuable information and can reflect the idea of overall tumor aggressiveness. However, our study has the following limitations: The Fractal computation has been applied on 2-D contour GTV regions, this study could be extended to 3D fractal analysis algorithm using a mesh approach, so that the region of GTV is better delineated for computation FD. Also, for a tumor contour which is very small, there might not be enough points in the number of boxes v/s the size of each box plane (log(N) vs log(1/r)) to describe the tumor contour. In this case, the best fit line will be only an approximation and hence this will be reflected in the FD value as well.

REFERENCES

1. *Freddie Bray et.al Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancer in 185 Countries, CA CANCER J CLIN 2018;0:1-31*
2. *Ma LH, Li G, Zhang HW, Wang ZY, Dang J, Zhang S, Yao L. The effect of nonsmall cell lung cancer histology on survival as measured by the graded prognostic assessment in patients with brain metastases treated by hypo-fractionated stereotactic radiotherapy. Radiat Oncol. 2016;11:92.*
3. *Yano M, Yoshida J, Koike T, Kameyama K, Shimamoto A, Nishio W, Yoshimoto K, Utsumi T, Matsumura Y, Moriyama S, Fujii Y. The outcomes of a limited resection for non small cell lung cancer based on differences in pathology. World J Surg. 2016. doi: 10.1007/s00268-016-3596-9.*
4. *Wu CC, Maher MM, Shepard JA. Complications of CT-guided percutaneous needle biopsy of the chest: prevention and management. AJR Am J Roentgenol. 2011;196(6):W678–W682.*
5. *Ravindra Patil et.al An Approach Toward Automatic Classification of Tumor Histopathology of Non–Small Cell. Tomography Jr.*
6. *Gillies R, Kinahan P, Hricak H (2016) Radiomics: images are more than pictures, they are data. Radiology 278:2*
7. *Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441–446.*
8. *Kumar V, Gu Y, Basu S et al (2012) Radiomics: the process and the challenges. Magn Resonance Imaging 30(9):1234–1248*
9. *Aerts HJ, Velazquez ER, Leijenaar RT et al (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative Radiomics approach. Nat Commun 5:40*
10. *Mandelbrot, B. B. The Fractal Geometry of Nature. New York: W. H. Freeman, 1982.*
11. *James W. Baish and Rakesh K. Jain, Fractals and Cancer, Perspectives in Cancer Research, 2000*
12. *Fractals: a possible new path to diagnose and cure cancer?, Future Oncol. (2015) 11(22), 3049–30513.*
13. *Heymans, O., Blacher, S., Brouers, F., and Pierard, G. E. Fractal quantification of microvasculature heterogeneity in cutaneous melanoma. Dermatology (Basel), 198: 212–217, 1999.*

14. *D. Cusumano et.al Development and Validation of New Radiomic Features Based on Fractal Analysis. International Journal of Radiation Oncology 2017.*
15. *https://www.wahl.org/fe/HTML_version/link/FE4W/c4.htm, last accessed and verified on 4th Nov 2018*
16. *Automated chest X-ray screening: Can lung region symmetry help detect pulmonary abnormalities?. IEEE Transactions on Medical Imaging. 2018*
17. *Feature Selection for Automatic Tuberculosis Screening in Frontal Chest Radiographs, Vajda, S., Karargyris, A., Jaeger, S. et al. J Med Syst (2018) 42: 146. https://doi.org/10.1007/s10916-018-0991-9*
18. *Foreign Circular Element Detection in Chest X-rays for Effective Automated Pulmonary Abnormality Screening. International Journal of Computer Vision and Image Processing(IJCVIP), IGI Global*
19. *Edge map analysis in Chest X-rays for Automatic Abnormality Screening, International Journal of Computer Assisted Radiology & Surgey (IJCARS), Springer*
20. *Szigeti, Szabó , Korom, Czibak, Horváth. Radiomics-based differentiation of lung disease models generated by polluted air based on X-ray computed tomography data. BMC Medical Imaging, Vol 16, Issue 1, Pages 14. February 11, 2016.https://doi.org/10.1186/s12880-016-0118-z*
21. *David Cusumano, Nicola Dinapoli, Luca Boldrini, Giuditta Chiloiro et al. Fractal-based radiomic approach to predict complete pathological response after chemo-radiotherapy in rectal cancer. La radiologia medica, 2017.*

CHAPTER 8

CLOUD BASED BIG DATA PLATFORM FOR IMAGE ANALYTICS

Sunil Kumar Vuppala
Dinesh M.S.
Sreramkumar Viswanathan
Ganesan Ramachandran
Nagaraju Bussa
Geetha Mahadevaiah

IEEE International Conference on Cloud Computing in Emerging Markets (CCEM) - 2017

ABSTRACT

This paper describes a framework for leveraging the cloud platform and big data technologies to perform medical image data analytics even across multiple institutions. The key challenges the authors have addressed are the security, privacy requirements, scalability of the solution and use of cloud based analytics for the existing proprietary legacy systems with network connectivity in hospitals. The Digital Imaging and Communication in Medicine (DICOM) medical images acquired for clinical purposes are to be shared with radiologists and institutions for interpretation of images to support disease diagnosis, who are miles away from the institutions. Prior to sharing DICOM images with third parties, the authors need to follow strict guidelines with respect to de-identification of Protected Health Information (PHI) from the medical data to be shared. To de-identify PHI information from images the authors have devised a newer generation de-identification techniques based on the big data platform to obscure privacy information from DICOM tags, embedded text in images and face information from CT and MR images.

Further to facilitate accurate analysis on medical images, a Clinical Decision Support System (CDSS) inference model is developed. In the paper, the authors proposed use cases and developed a cloud based framework to distribute/share clinical images with CDSS to radiologists on the move and to the institutions that do not exists in the host organization premises. The developed solution satisfies all the PHI de-identification requirements and supports scalability by efficiently using big data cloud technologies, for faster and reliable processing. The Gross Tumor Volume (GTV) application using deep learning approachis an example to illustrate CDSS.

Keywords

DE IDENTIFICATION
PICTURE ARCHIVING AND COMMUNICATION SYSTEM
MEDICAL IMAGE ANALYTICS
BIG DATA PLATFORM
DEEP LEARNING IN CLOUD

1.0 INTRODUCTION

The Digital Imaging and Communication in Medicine (DICOM) is the standard for storage, viewing and transmitting medical imaging data. Medical images acquired for clinical purposes are shared with third parties such as research institutes, radiologists and teaching institutions. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy rule specifies strict guidelines with respect to de-identification of Protected Health Information (PHI) from the medical data to be shared. DICOM anonymizer refers to a tool that removes or de-identifies Protected Health Information (PHI) of the DICOM images prior to sharing DICOM images with third parties.

The advances in diagnostic imaging technologies and high volume, longitudinal examinations using imaging data acquired through a multitude of modalities have put enormous pressure on Healthcare enterprises that has given rise to a storage crisis. Vendor neutral secured storage enabled cloud based picture archiving and communication system (PACS) solution is proposed to archive de-identified DICOM images in cloud.

In the present hospital environment, there exists legacy modalities such as Ultrasound, CT and MRI machines do not have the capability to connect seamlessly to the cloud and apply machine/deep learning methods. Our proposed platform handles this gap of cloud connectivity to carry out analytics in a scalable way.

Further, to analyze the stored DICOM images, the authors deployed deep learning pre-trained models on a cloud based Picture Archiving and Communication system (PACS) with AI inference engine, to identify the organ of interest in CT images.

The theme is to acquire, anonymize, store in cloud and perform image analytics. The contributions in the paper include:

- A framework for medical image analytics using cloud based big data platform
- The framework includes de-identification, vendor neutral security enabled PACS system on cloud which acts as place for performing CDSS applications
- Proposed various use cases for the framework

• Enabling connectivity for legacy medical devices that can push the data to cloud and obtain analysis

• A deep learning based ultrasound scenario implementation.

The rest of the paper is organized as per the following sections. Section II outlines an overview of existing de-identification techniques and challenges, available PACS systems and deep learning reference implementation details. The framework of proposed cloud based big data platform for medical image analytics is discussed in section III. Various use cases using the proposed framework are discussed in section IV. Modeling of ultrasound DICOM images to identify the organ of interest in obstetrics is discussed in section V. The experimental details are presented and results are analyzed in section VI. Section VII contains conclusion and future directions.

1.0 LITERATURE SURVEY

Anonymization Needs and Challenges

Data sharing is increasingly recognized as critical to cross-disciplinary research and to assure scientific validity. Despite National Institutes of Health and National Science Foundation policies encouraging data sharing by grantees, little sharing of clinical data has in fact occurred. A principal reason often given is the potential of inadvertent violation of the Health Insurance Portability and Accountability Act privacy regulations [1]. While regulations specify the components of private health information that should be protected, there are no commonly accepted methods to de-identify clinical data objects such as images. This leads institutions to take conservative risk-averse positions on data sharing. In imaging trials, where images are coded according to the DICOM standard, and the complexity of the data objects and the flexibility of the DICOM standard have made it especially difficult to meet privacy protection objectives. The recent release of DICOM Supplement 142 on image de-identification has removed much of this impediment [2].

The field of header tag de-identification is a well-matured field and many third party tools provide support for anonymization of DICOM tags. A number of toolkits exist for DICOM header tags de-identification. A comprehensive list of DICOM anonymizers along with their download locations is provided at [3]. A recent study comparing ten DICOM anonymizer tools is reported in [4].

However, most of these tools do not have capabilities to de-identify PHI text information that are embedded in DICOM images ('burned-in' PHI text). It is also evident from literature that few tools and some researchers have tried to do face recognition from surface or volume rendered CT and MR images and they were able to achieve moderate success [5-6]. In order to avoid leaking of personal face and PHI text information to outside world while sharing research data it is also required to mask PHI text and face information from DICOM images.

During our literature study, no single tool was found that could provide all three types of de-identification functionalities together. The evaluation of selected tools show that the field of header tags de-identification has matured and many of the tools provide support for encryption/decryption of tags information. The support for embedded text de-identification is mainly semi-automatic, where the user has to manually configure the coordinates of the rectangle to be blackened [5] or open the file in the editor and perform manual selection of the rectangle to be blackened [6]. The facial mask de-identification tool such as MBIRN defacer [7] are automatic, however they are not generic enough to de-identify all possible DICOM volumes containing face.

Big Data Platforms for Image Analytics

Most of the PACS systems in hospitals have the capabilities to store and visualize medical images. Some of the hospitals rely on a cloud based PACS that stores and backs up data in an offsite server and not within the organization's physical location. Existing PACS mainly focus on storage, user authentication, backup, and visualization. Usually they are not intelligent due to a lack of AI models deployment and they are not scalable in nature. The authors propose to extend the capabilities of PACS with an AI inference engine.

The advancement in clinical & medical imaging and the rapid rise of data during the course of medical procedures has brought about a need for big storage and archival mechanisms. Multi-vendor ecosystems further complicates image acquisition. Detailed imaging allows radiologists and clinicians to look at things in a new perspective and can provide more accurate and timely diagnosis [8].

Historically, all legacy PACS have been an on premise and closed solution. The issue with this approach is that the image sharing across health care enterprises is almost impossible [9].

However, with cloud technology advancements and de-identification modules being embedded and widely used, it is imperative to move the PACS also to the cloud and many PACS vendors have already forayed into that path. Nevertheless, most of them either provide cloud-only or premise-only models. However, the need of the hour is go with a hybrid model [11] where the solution works both on premise and extend to cloud model on-demand either for archival or for access across health care professionals and data scientists.

Another aspect about PACS systems has been the storage of standardized DICOM images, which is helpful for data interoperability, but additionally the system has to extend to save non-DICOM images giving true vendor neutrality and more so in a multi-vendor ecosystem [10]. Cloud storage also enables models like pay-per-storage/ pay-per-access for healthcare organizations help reduce Total cost of ownership (TCO).

Finally, the huge haystack of images acquired at scale needs to be put to meaningful use. One way to use this data is to analyze in real-time or in near real time to screen and treat patients with for example Pulmonary tuberculosis using chest X-ray [12].

There are many other scenarios like tumor classification, cohort clustering based on similar medical conditions, medical record associations, predictions, decision making and many more that would benefit from access to this data[13] .

Clinical decision support algorithms are evolving over time from simple decision tress (rule based) to deep neural networks (DNN). With advancements in algorithms, requirement on infrastructure to run them in real time have also shot up. Figure 1 shows the number of operations to be performed in DNN architecture in the recent years [14].

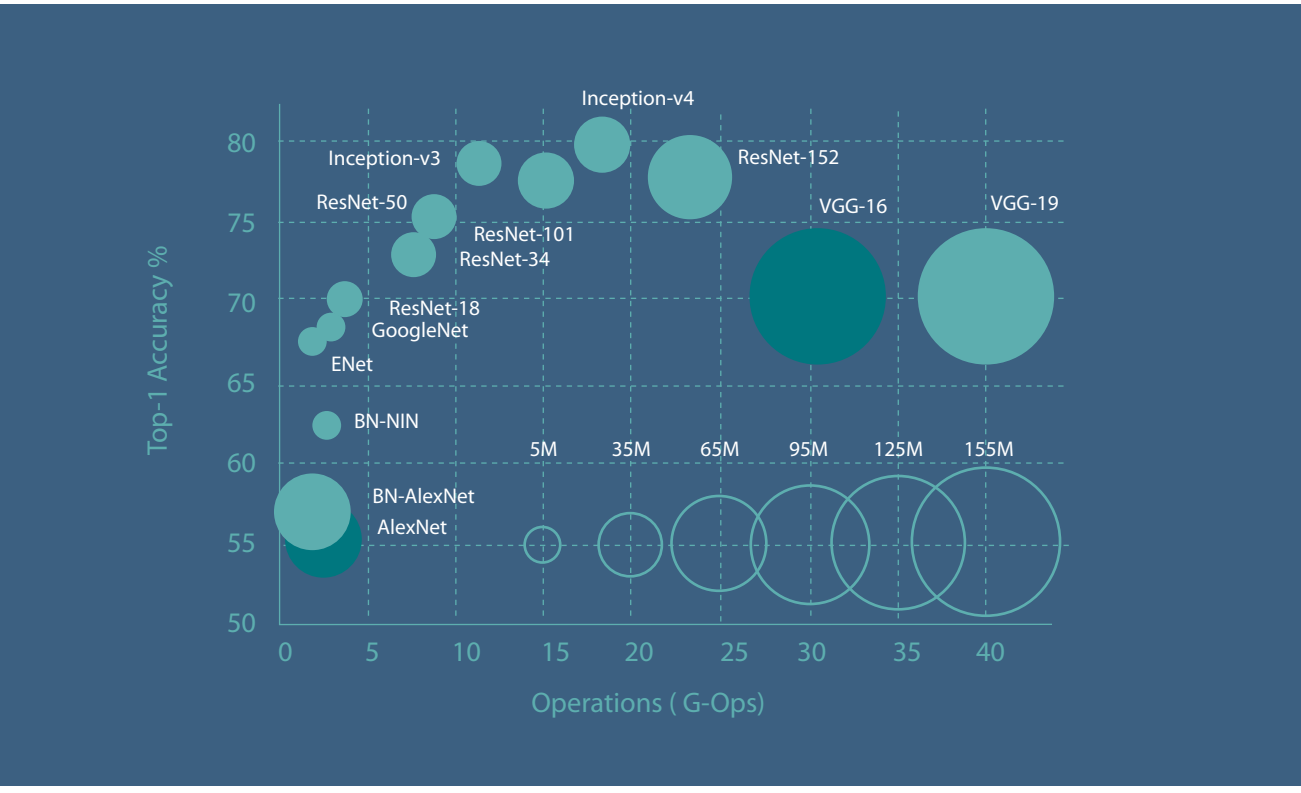


FIGURE 1: ***Number of operations for standard networks***

While the cost of GFlops (Giga floating-point operations) have consistently decreased[15] as shown in Figure 2, the demand for higher computational power has made hardware obsolete in a shorter time.

FIGURE 2: ***Cost of GFlops over time***

Since the infrastructure needs to cater to different complex models (from identification to segmentation), the infrastructure should be scalable as required and shareable for cost efficiency.

3.0 FRAMEWORK

In an Imaging workflow, medical devices such as X-ray, CT, MR, Ultrasound can acquire patient images in a noninvasive way to reveal the information on the internal organs. Selection of the imaging modality depends on the type of ailment and doctors/ radiologist choice. In this framework shown in Figure 3, the authors have devised a three-stage cloud based big data platform solution for image analytics. In the first stage, a PHI anonymization tool is devised to de-identify PHI information that exists in text and images. Second stage, on the images uploaded to cloud, lesion or organ of interest marks will be generated on images with annotations. Generation of lesion annotations is performed based on the Clinical Decision Support system (CDSS) Inference Engine deployed on the cloud PACS. Third stage, based on the user requirement, images along with lesion marks will be delivered to Radiologist's devices for final adjudication on the CDSS marks. The need for a big data platform is to address the needs of scalability and to handle online and offline analytics in cloud.

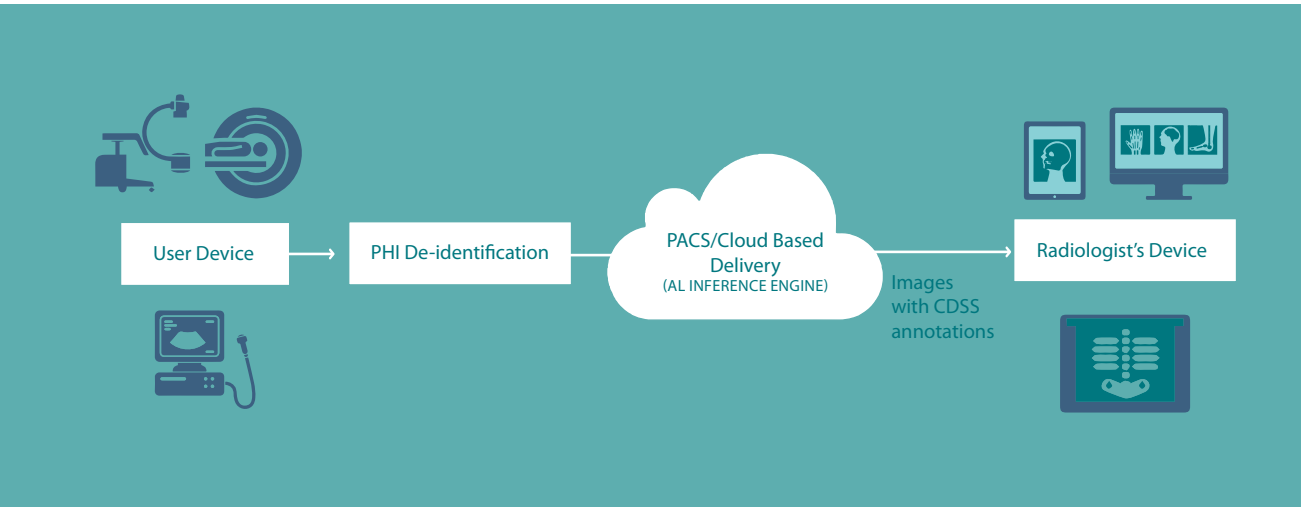


*FIGURE 3**: Framework for cloud based big data platform***

DICOM De-identification using Big Data Platform

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy rule specifies strict guidelines with respect to de-identification of Protected Health Information (PHI) from the medical data to be shared. A big data platform for DICOM PHI anonymizer was developed.

In DICOM, a Data Element Tag uniquely identifies a Data Element. The DICOM standard defines UIDs for describing the hierarchy from top to bottom as follows:

- Study UID - Identifier of the study or scanning session
- Series UID - The identifier within a series acquired in one scan
- Image UID – The identifier which should be unique for any image

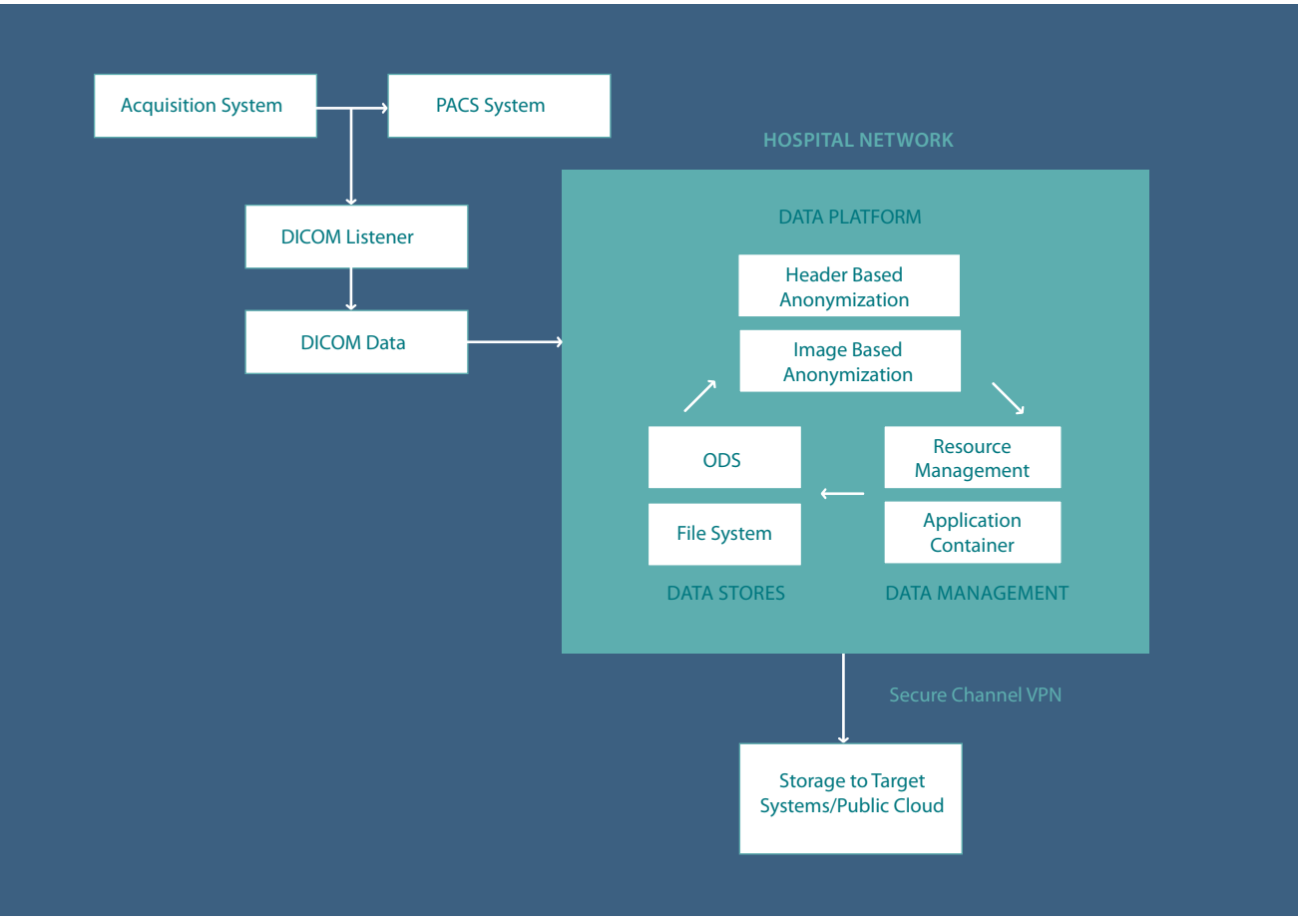


FIGURE 4: ***DICOM PHI Anonymizer***

The overall architecture of the proposed DICOM anonymizer is shown in Figure 4.According to this architecture, the DICOM listener component intercepts and collects DICOM images flowing from the DICOM acquisition system (CT, MRI, ultrasound, angiography or fluoroscopy modalities, etc.,) to the Picture Archiving and Communication System (PACS). The anonymization functions used in the DICOM anonymizer includes DICOM header (Tags) de-identification functions and embedded text masking function. The images retrieved by the DICOM Listener are stored as a DICOM volume in a file system. The Anonymization tool takes as input DICOM images from the file system. These images undergo the functionality of (i) Header Based Anonymization: DICOM header (Tags) anonymization and (ii) Image Based Anonymization: embedded text de-identification and face de-identification respectively.

Resource management refers to the functionality of assigning resources for running the DICOM anonymizer in a distributed environment, while the application container is responsible for storage and management of the DICOM anonymizer application (e.g. executable java application) on the hospital premises which stores mapping information of the original v/s the anonymized patient data. The de-identified images are sent over the secure channel to the target systems or public group as per the requirement.

The big data platform collects and processes data from the primary data sources in which data is first generated and then outputs the final analysis results using storage, analysis and search processing. The output analysis is usually shown visually using graphs and/or charts.

Figure 5 shows an example of a de-identified ultrasound image with embedded text and application of embedded text mask.

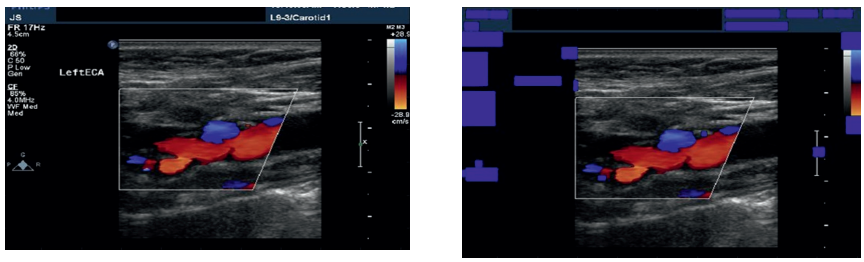


FIGURE 5. ***Ultrasound image with Embedded Text and Embedded Text Mask***

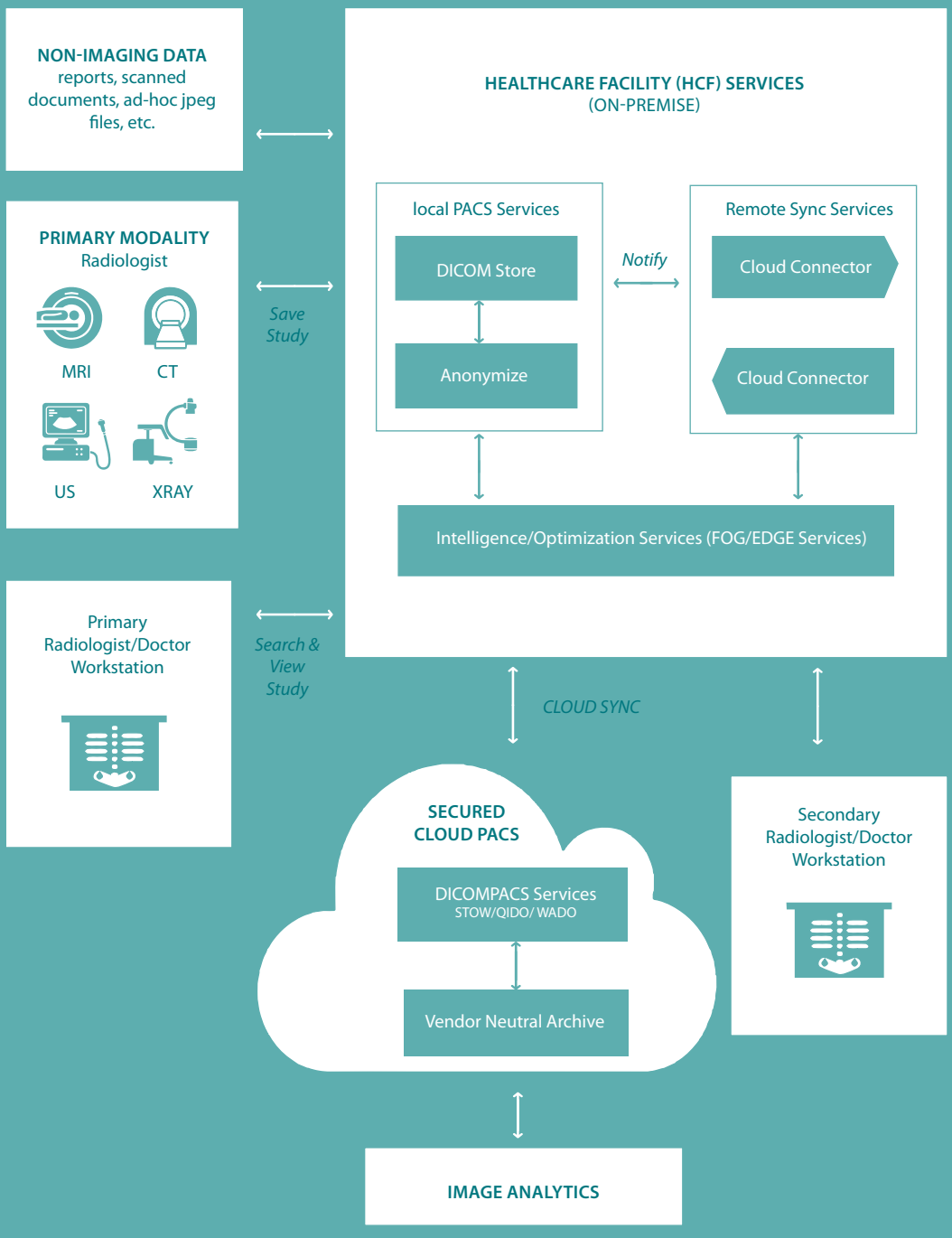


FIGURE 6. ***Vendor Neutral Secured Storage Enabled PACS System on Cloud***

The Figure 6 depicts the cloud based PACS eco-system.

• The Radiologist performs the scan on the subject and from his workstation saves the study into the local PACS system.

• As the scans enter the system, they are anonymized/de-identified and notified to a cloud connector.

The cloud connector ingests the anonymized/de-identified data into a secured cloud-based Vendor-neutral PACS platform. Now with the study available in the system, the Artificial Intelligence inference engine (model) deployed in PACS can be applied on the images that are available in PACS.

Table 1 covers the software components used in Big Data Platform. With the use of this platform, users can also analyze hospital department key performance indicators in system utilization and operational efficiency space.

SOFTWARE COMPONENT	DESCRIPTION
Hadoop/HDFS	Distributed File System for storing input/output images
MapR-DB /No SQL	Schema Less Store of DICOM files
Mesosphere	Resource Manager
Apache Spark	Processing Engine
Java	Anonymization functions
Finagle	Creating client SDK
Hashing with Salt	MD5 algorithms
Apache Commons	Supporting multiple distribution systems (HDFS, S3 etc.) - in progress
Python	Image processing Algorithms
Scala	Programming language for anonymization APIs
SBT	Build tool for packaging anonymization jar
Junit	Testing framework for anonymization functions
ScalaTest	Testing framework for anonymization APIs

*TABLE 1. S**oftware components used in big data platform of DICOM anonymizer***

Picture Archiving and Communication System (PACS) in Cloud

The advances in digital imaging technologies demand a lot of storage in healthcare enterprises. Supplementary to this, data de-duplication, enrichment and contextualization that helps in diagnostic precision, brings the need for linking intelligence and meta-data around imaging. Furthermore, the miscellany of econometrical applications on diagnostic imaging has given rise to varieties of interesting use cases where storage and sharing of imaging data is of utmost importance.

Over and above, security plays a very important role in ensuring data privacy. As a consequence, Vendor Neutral Secure Storage Enabled Enterprise PACS system on cloud is required to ensure that the on-demand, interoperable and ubiquitous data sharing is seamless across the continuum of care. It also ensures high availability of data, whenever and wherever needed. Meta-data about the ingested studies could utilize an unstructured data store for fine-grained search and access.

The security aspects of data access are managed by role based user authentication, including audit trails as per US HIPPA Standards. Thus, the data on the vendor neutral cloud based PACS might play a role in making data more FAIR, which is Findable, Interoperable Accessible and Re-usable.

Proposed on cloud storage also provides capabilities to store non-imaging data like investigative reports associated to an imaging study, or doctors' scanned assessment reports and the like. Intelligence layers could also provide opportunity to leverage hybrid models in PACS systems.

Additionally, one could leverage cross-reference of patient lookup information to synchronize data between on cloud and on premise medical systems.

Deliver images with CDSS marks on user devices

Based on the user requirements, images along with lesion marks will be delivered to radiologist's devices for final adjudication on the CDSS marks as depicted in the use case diagram in Figure 7.

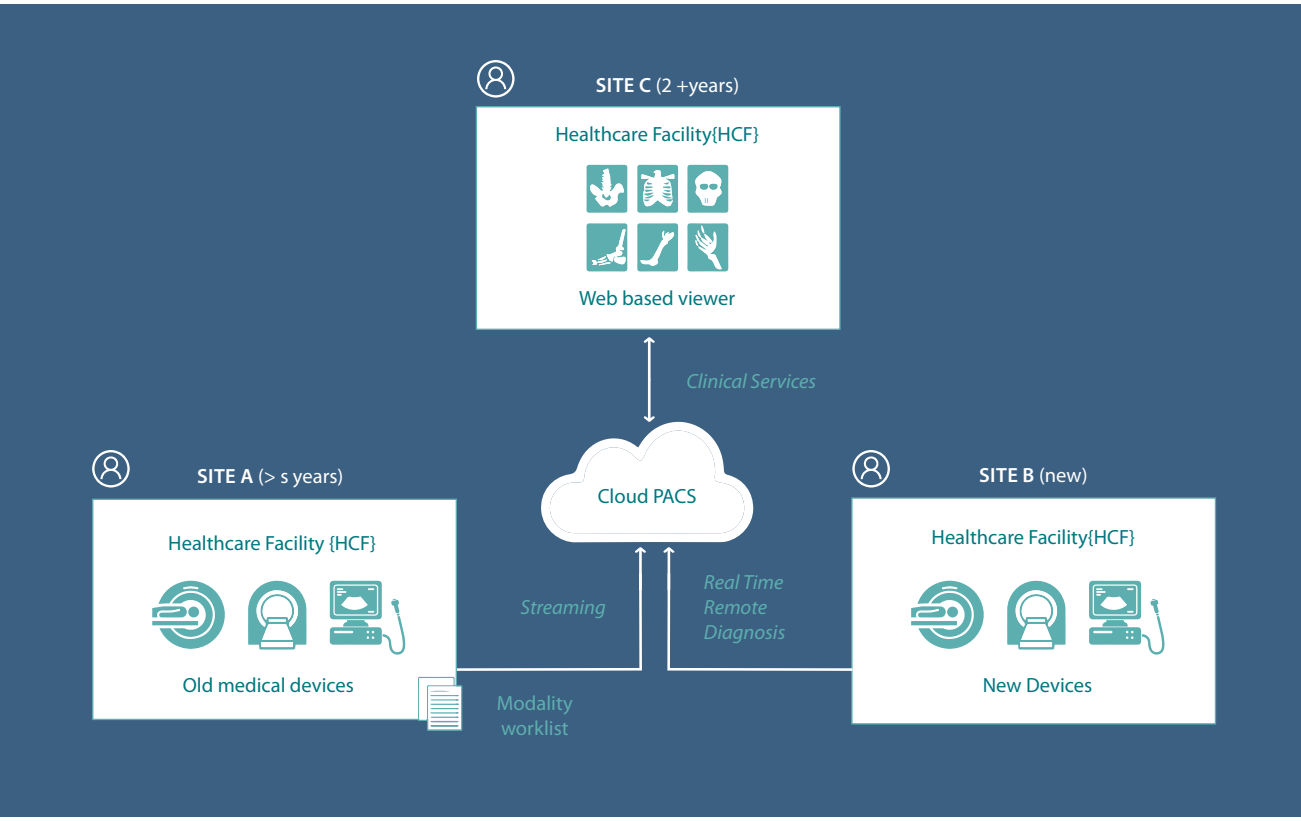


FIGURE 7. ***Image Acquisition, Sharing and Analytics***

Using web based access, user devices can pull images along with CDSS marks from PACS for final reads. Various research studies confirms that reading images along with CDSS marks enables better image interpretation with increased accuracy[17].

USE CASES using the proposed framework

The authors discuss the following use cases for the proposed framework.

USE CASE 1 - Chain of hospitals with multi-vendor legacy devices: In this use case, a group or chain of hospitals may have many medical devices procured at different instances of time. Our focus is to extend the capability of legacy systems with new deep learning based algorithms in a scalable manner. For instance, if a radiologist wishes to use a new deep learning based tumor detection on a 5 years old MRI machine of Philips, then radiologists can seamlessly use the proposed framework.

USE CASE 2 - University setup / Research organizations: This use case is for multiple clinical studies planned at different clinical sites. Data may need to be collected from multiple sources to train and evaluate research questions. Later the same can be extended to a product or large scale solution. The proposed framework can be used in both the scenarios.

USE CASE 3 - Diagnostic centers/ Polyclinics: This use case covers the small establishments such as diagnostic centers, which may have very few (1-10) medical devices. In this use case, with the use of the proposed framework, these organizations can save maintenance of infrastructure and radiologists cost on premises.

USE CASE 4 - Real time screening: In this use case, the framework can be used to carry out real time analytics on medical images. The only requirement here is to have good network connectivity from the location to the cloud so as to stream the data to the cloud and get diagnostic inference. Artificial Intelligence model training is taken offline and the trained model is deployed in the cloud for driving inference. Example scenarios include screening at the airports and malls for specific ailments using x-ray images.

USE CASE 5 - Tele medicine: Another interesting use case is telemedicine. Especially in developing countries, there are many villages/towns where there is a scarcity of clinicians. So, the framework can be used to add analysis on top of the diagnosis image so that the radiologist can analyze them remotely in a faster pace.

4.0 MODELING

In this section, the authors present Deep Learning (DL) model to demonstrate the application of the framework. Machine learning and deep learning technologies are gaining popularity in medical image analytics due to the availability of big data platforms. The clinical decision support system based on the image analytics on big data, can lead to improved productivity and accuracy of diagnosis by radiologists and clinicians. The authors designed a system using U-net deep neural network to automatically delineate the Gross Tumor Volume (GTV) in CT Lung studies. Architectural details of the system are given in Figure 8.

As a first step, anonymization of data was done, before deep learning technique was applied on the CT images on the big data platform.

The challenges of segmenting a part of the lung as a 'Region of Interest' in CT images and delineating GTV, is complicated by inter-observer variability, object variability, image quality and varying tumor sizes .

We used the platform to detect a GTV using the deep learning model as shown in Figure 8.

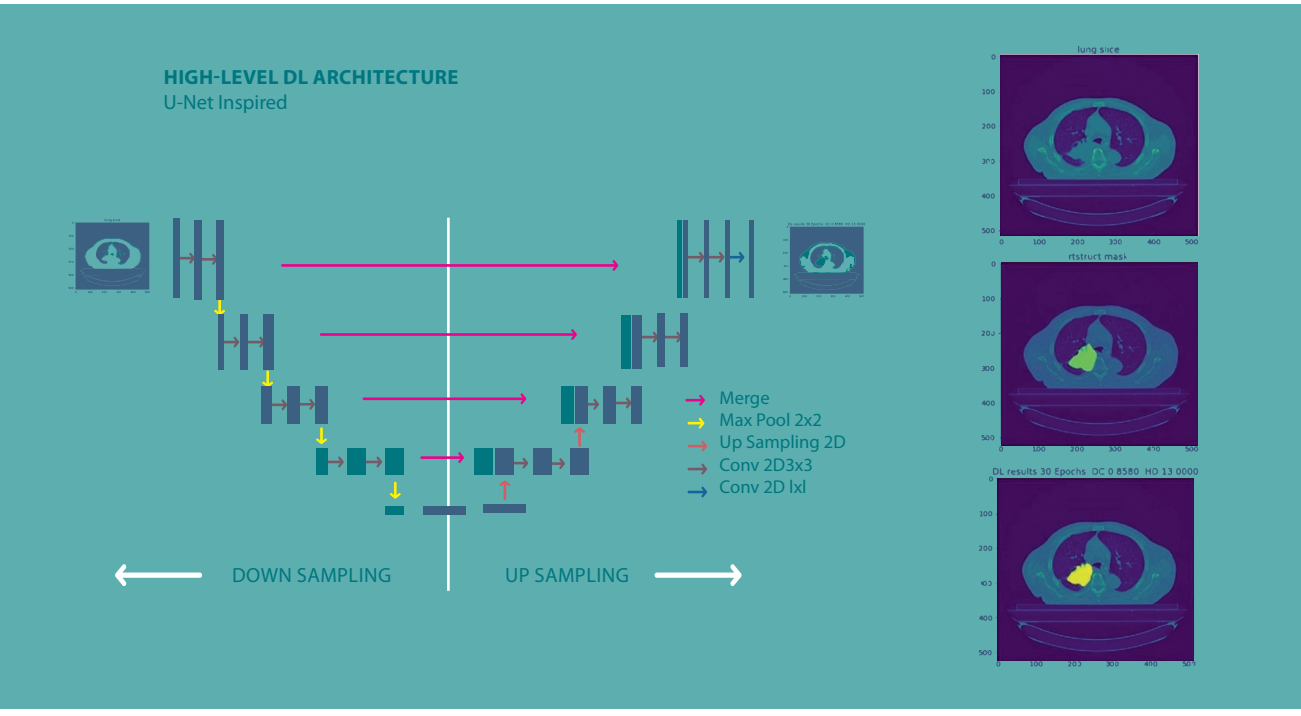


FIGURE 8. ***DL model architecture and Gross Tumor Volume results***

5.0 EXPERIMENTS & RESULTS

The authors implemented the proposed model in a cloud environment. The goal was to find the latency and requirements of network speed to have real time performance. Here are the observations from experimental setup:

It is observed that the cost of data maintenance 'on-premise' is more than data transfer to the cloud. Sharing the results among the doctors for a secondary opinion is an additional benefit by using this platform instead of an 'on-premise' solution. For instance, a 100Mbps in Bangalore (As on July 2017, with Rs 64/$) costs Rs 8/GB ($0.12/

GB). A typical CT scan can go up to 300MB, which means the transmission and reception cost of such scan would be 4 cents. Cost of storing the data and accessing in the cloud is 3 cents/GB/month [16] whereas the on premise the cost is 3x (9 cents/GB/month) including the backup and maintenance.

For CT scans, the collection time is typically 10 mins for 200 frames with each frame of size 512x512. If demand for results is one minute then a 10 Mbps connection is sufficient. The processing time in the cloud is micro to milliseconds for inference based on the model and the infrastructure.

6.0 CONCLUSIONS

In this work, the authors have developed a framework to anonymize DICOM images using a big data platform, store de-identified images in cloud and perform image analytics on the acquired images. Once the data is acquired from the medical device, a DICOM anonymizer tool is used to automate tags de-identification and embedded text masking. The data is stored in cloud based on a PACS eco-system with CDSS capability. The proposed framework with CDSS enables better interpretation of images with increased accuracy. As the framework is big data based there is built-in support for scalability. This framework also helps in extending the legacy systems to use cloud based analytics. As an example, the authors have applied deep learning on images to identify Gross Tumor Volume(GTV) images in the cloud based big data platform. The same image analysis can be extended to signals such as ECG from patient monitoring and Ultrasound scans as well.

REFERENCES

1. *John B. Freymann et. Al, "Image Data Sharing for Biomedical Research—Meeting HIPAA Requirements for De-identification" J Digit Imaging (2012) 25:14–24.*
2. *David Clunie, De-identification Revisited DICOM Supplement 142, DICOM INTERNATIONAL CONFERENCE & SEMINAR, Oct 9-11, 2010 Rio de Janeiro, Brazil.*
3. *DICOM Anonymizer Software, available online: http://dicom-anonymizer.winsite.com/, Last accessed date: 08th Nov. 2016*

4. *Aryanto, K.Y.E., Oudkerk, M. and van Ooijen, P.M.A., 2015. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. European radiology, 25(12), pp.3685-3695*

5. *RSNA MIRC Clinical Trials Processor (CTP), available online: http://mircwiki.rsna.org/index.php?title=CTP-The_RSNA_Clinical_Trial_Processor, Last accessed date: 08th Nov. 2016.*

6. *Bischoff- Grethe, A., Ozyurt, I.B., Busa, E., Quinn, B.T., Fennema Notestine, C., Clark, C.P., Morris, S., Bondi, M.W., Jernigan, T.L., Dale, A.M. and Brown, G.G., 2007. A technique for the deidentification of structural brain MR images. Human brain mapping, 28(9), pp.892-903.*

7. *Mri_deface: Automated Defacing Tools, Available online: https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface, Last accessed date: 13 Nov. 2016.*

8. *Imaging Adavnces: http://www.webmd.com/a-to-z-guides/features/diagnostic-imaging-beam-me-up-dr-mccoy#1, Last Accessed date:26 July 2017.*

9. *Cloud PACS: http://appliedradiology.com/articles/roadmap-to-cloud-based-pacs, Last Accessed date26 July 2017.*

10. *Vendor Neutrality: https://www.itnonline.com/article/what-vendor-neutral-anyway, Last Accessed Date 26 July 2017*

11. *Hybrid Models: https://www.itnonline.com/article/what-vendor-neutral-anyway, Last Accessed Date 26 July 2017.*

12. *Image Analytics: http://www.carestream.com/blog/2016/05/23/what-is-future-of-big-data-in-radiology/, Last accessed Date 26July 26, 2017*

13. *Image Mining: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4094980/, Last Accessed Date 26 July 2017.*

14. *Alfredo Canziani, Adam Paszke, Eugenio Culurciello, "An Analysis of Deep Neural Network Models for Practical Applications", CoRR, abs/1605.07678, 2016*

15. *Trends in the cost of computing Available online: http://aiimpacts.org/trends-in-the-cost-of-computing/ Last accessed date: 20 Jun 2017*

16. *Amazon S3 Pricing, Available online: https://aws.amazon.com/s3/pricing/ Last accessed date: 30 Jun. 2017*

17. *Karami, Mahtab. (2015). Clinical Decision Support Systems and Medical Imaging. Radiology management. 37*

CHAPTER 9

MACHINE AND DEEP LEARNING BASED CLINICAL DECISION SUPPORT IN MODERN MEDICAL PHYSICS: SELECTION, ACCEPTANCE, COMMISSIONING AND QUALITY ASSURANCE

Geetha Mahadevaiah
RV Prasad
Inigo Bermejo
David Jaffray
Andre Dekker
Leonard Wee

Journal : Medical Physics, 10.1002/mp.13562, May 2019

ABSTRACT

Recent advances in machine and deep learning based on an increased availability of clinical data have fueled renewed interest in computerized clinical decision support systems (CDSSs). CDSSs have shown great potential to improve healthcare, increase patient safety and reduce costs. However, the use of CDSSs is not without pitfalls, as an inadequate or faulty CDSS can potentially deteriorate the quality of healthcare and put patients at risk. In addition, the adoption of a CDSS might fail because its intended users ignore the output of the CDSS due to lack of trust, relevancy or actionability. In this article, we provide guidance based on literature for the different aspects involved in the adoption of a CDSS with a special focus on machine and deep learning based systems: selection, acceptance testing, commissioning, implementation and quality assurance. A rigorous selection process will help identify the CDSS that best fits the preferences and requirements of the local site. Acceptance testing will make sure that the selected CDSS fulfills the defined specifications and satisfies the safety requirements. The commissioning process will prepare the CDSS for safe clinical use at the local site. An effective implementation phase should result in an orderly roll out of the CDSS to the well-trained end-users whose expectations have been managed. And finally, quality assurance will make sure that the performance of the CDSS is maintained and that any issues are promptly identified and solved. We conclude that a systematic approach to the adoption of a CDSS will help avoid pitfalls, improve patient safety and increase the chances of success.

Acknowledgements

The authors acknowledge support from projects BIONIC and STRaTegy, funded by the Netherlands Organisation for Scientific Research (NWO), as well as structural funding for LW and AD.

1.0 INTRODUCTION

The recent prominence of artificial intelligence (AI) and machine learning (ML), coupled with the growing volume of available clinical data, has led to an increased interest in applications of AI in general [1] and of computerized clinical decision support systems (CDSS) in particular. A computerized CDSS is any software designed to aid clinicians and patients in clinical decision-making, defined as "active knowledge systems which use two or more items of patient data to generate case-specific advice", according to Wyatt and Spiegelhalter [2]. CDSSs can make use of expert knowledge and/or models learnt using statistics and ML from data.

In the early days of CDSSs, they were conceived as being able to eventually replace the clinician's decision-making. A nuanced, more modern view of the purpose of CDSSs is to assist the clinician to make better decisions than either the clinician or the CDSS could make on their own, by processing the vast amount of available information.

Typically, a modern CDSS makes recommendations to the clinician, and the clinicians are expected make their own decisions and over-ruling CDSS recommendations they believe to be inappropriate. Computerized CDSS have evolved dramatically since their first steps featuring the computer aided diagnosis in the Leeds Abdominal Pain system[3], the rule based MYCIN [4], and the HELP alert system [5]. One way they have evolved in is their integration into clinical workflows and other clinical information systems: in the beginning, they were standalone systems where clinicians had to enter the patient information before reading and interpreting the results. Beginning in 1967, CDSSs started to be integrated into clinical information management systems thus offering two main advantages: users did not have to re-enter information, and CDSSs could be proactive, i.e. alerting or recommending actions, without the user actively seeking assistance from the CDSS [6]. Starting in the late 1980s, the development and adoption of standards to represent, store and share clinical knowledge allowed separation of knowledge content from the software code of the CDSS [7]. From 2005, clinical information systems started offering application programming interfaces (APIs) through which they could interact with CDSS, thus allowing for a more dynamic and less standardized relationship [8].

The evolution of CDSSs has led to a high variety of CDSS types [9] , which can be classified in terms of a number of features. CDSS can offer support on demand or unprompted, as is the case of alert systems.[10] In addition, CDSSs can be classified in terms of their underlying technology as based on rules, deep learning [11], probabilistic models, genetic algorithms, or reinforcement learning[12], amongst others. In terms of their function, CDSSs can be classified as supporting diagnosis, outcome prediction [13]. treatment planning [14], prescribing and managing medications,[15,16] preventative care [17], chronic disease management [18], image interpretation (contouring [19], segmentation, and pathology detection), and many others.

Systematic reviews suggest that use of CDSSs reduces unwarranted practice variation, improves quality of health care, reduces waste in the healthcare system and reduces the risk of overload and burnout among clinicians.[20-24] However, CDSSs can also have important negative consequences, since a faulty CDSS or its inappropriate use can lead to deterioration of the quality of care. Major ethical questions and patient safety concerns still remain.[25] With the advent of deep learning, CDSSs are becoming better and as they reach human performance levels, are showing behavior indistinguishable from humans.[26] This raises new questions regarding responsibility and liabilities; however, these are outside the scope of the current review. Regulatory processes are already in place to alleviate some of these concerns, and they keep evolving in order to keep their currency in this rapidly developing field.[27] However, a regulatory approval is not a guarantee of good performance. CDSS can inadvertently increase the workload of the clinicians. For example, a well-known consequence of a CDSS alerting system in patient monitoring is "alert fatigue", that occurs when clinicians come to ignore alerts due to an overwhelming frequency of false alarms. [28] Another potential risk arising from the adoption of CDSSs is clinicians losing the ability to make decisions on their own or to determine when it is appropriate to override the CDSS – and again current gains in artificial intelligence, which make it a reality that CDSS are equal or better in decisions making than humans, make these risks more pertinent. This could become critical in case of computer system downtime, or if a patient with an unusually rare medical condition is admitted for treatment. As such, it is important to remain alert to both the positive and negative potential impact of CDSS on clinical decision making.[22] Some forms of CDSSs have been in use for decades, but their use is not yet widespread due to a number of issues related to design and implementation, such as clinicians not using them due to lack of time or lack of confidence in the CDSS's output.[29,30]

However, there remains an immense potential need for CDSSs due to increasing volume of available data, growing diversity of treatment options and rapidly evolving medical technologies. CDSSs could be valuable as a means of delivering medical care tailored towards patients' preferences and biological characteristics. Patients could benefit from an overall accumulation of human knowledge and clinical expertise guiding their diagnosis, treatment and condition monitoring. There remains a growing global need for high-quality personalized medicine to improve patient outcomes, reduce financial burden and avoid unwarranted practice deviations. Machine learning based CDSSs are expect to help alleviate some of the current knowledge and associated quality of care variation across countries and regions. Thus, the question of designing, developing, presenting, implementing, evaluating, and maintaining all types of clinical decision support capabilities for clinicians, patients and consumers remains a key area of research in modern medicine.[31]

The aim of this paper is to provide guidance to select, test, commission, implement and maintain CDSSs in a clinic safely and successfully. The paper is organized as follows: the next section explains how to select a CDSS; the next to section provide recommendations for the acceptance testing and commissioning of a CDSS; then, the implementation section describes how to roll out a CDSS while section 6 provides guidelines for the quality assurance of CDSSs; finally, we draw some conclusions.

2.0 SELECTION

The range of commercially available CDSSs for clinical applications has been growing during recent years. Hence, selecting the most appropriate CDSS from those available is not always easy yet it is a key step in the implementation of a successful CDSS.

User acceptance of CDSS is critical: several implementation studies [32,33] show that how beneficial a CDSS is perceived largely determines uptake and usage by clinicians and allied health professionals. Therefore, the recommended first step in the process would be to form a multi-disciplinary steering committee comprising key clinical stakeholders, such as a number of clinician "champions", patient representatives, department administrators and information technology experts, who would be willing to take decisions and be accountable for the implementation of a CDSS.[34] Studies show that likelihood of user acceptance increases when CDSS implementation involves the end users instead of being forcing the CDSS onto the end users.[32] In order for it to be effective, the CDSS should be conceived as part of a wider, coherent and department-wide quality improvement strategy, where a clinical quality gap between current patient outcomes or process and the desired end-state has been clearly identified and carefully measured. [35]

Two main aspects to consider when selecting a CDSS are the quality of the CDSS and how well the CDSS fits with closing the clinical quality gap. The quality of a CDSS needs to be considered at least at two levels: the level of the technology platform and that of the data or knowledge used to build it. CDSSs, as software that is potentially also a medical device, should be designed, implemented, tested and documented using generally recognized quality assurance methods for software development used in the medical domain. The medical knowledge used in the construction of the CDSS cannot be proven clinically complete or objectively correct, but it must attempt to capture the current state of professional and scientific opinion. Further, it must be possible to verify formally that the relevant medical knowledge satisfies certain requirements such as being unbiased, consistently interpreted and reasonably complete.[36] In the case of CDSS based on models learnt using statistical analysis or by machine learning, an assessment of the quality of the source data is necessary. Data quality is important, since the "garbage in, garbage out" principle especially applies to machine learning. Data is generally defined as of high quality if it fits closely to the intended purpose,

[37] and more specifically it should consist of a representative, unbiased sample of the domain (patients, or clinical conditions) being modeled. The appropriate processes for anomaly detection, data cleansing, handling of incomplete or missing data should have been applied to the dataset, and the existence of potential biases assessed and corrected.

A key indicator of the quality of a CDSS is its performance. Measures of performance vary across different types of CDSS. For example, in CDSS performing outcome prediction, the area under the receiver operating characteristic (ROC) curve or the c-index are commonly used performance metrics.[13] In other cases, performance can be measured in terms of saved time.[19] However, the assessment of the performance might be complicated, [38] especially when a gold standard of performance does not exist, such as in the case of therapy-advice systems, where even experts may disagree. In the end, the most difficult to measure, yet most valuable performance metric, is the effect of the CDSS on health outcomes or processes.[35] Publication by CDSS vendors of detailed evaluations of usability and effectiveness of CDSS implementation might facilitate purchasing decisions, [34] but it should be kept in mind that trials conducted by developers of CDSS might overestimate their benefits and third party external validation is required. [39] A thorough hazard analysis, resulting in an exhaustive list of potential risks and their possible consequences along with a mitigation plan for said risks, [36] is part of the regulatory process and could provide valuable insights into the desirability of the CDSSs.

During selection, the acceptability of the CDSS should be considered and weighed against performance. For users to easier accept the output of a CDSS, the strength of evidence supporting the clinical recommendations delivered by the CDSS should be transparent to the user [40] The levels of comprehensibility or explainability of models based on hand-engineered features and simple models (e.g. decision trees) is usually higher than those based on more advanced approaches such as random forests and deep learning. [41]

As mentioned earlier, it is crucial to select a CDSS that fits the requirements of the local site. First, following the Population, Intervention, Comparison and Outcome (PICO) framework, [42] the selection process should be restricted to CDSSs that target the appropriate population, consider the relevant intervention and comparators and focus on the outcomes of interest. When selecting a CDSS, we should consider the five

"rights" a CDSS should fulfil, namely: delivering the right information (what), to the right people (who), in the right format (how) through the right channels (where) at the right time in the workflow (when). [43] Delivering the right information also implies that the output of the CDSS (clinical recommendations and assessments) should be clinically relevant, brief, unambiguous and actionable. [40] The CDSS should also fit the existing workflow of its users as closely as possible, e.g. integrated in the electronic health record (EHR), minimizing the effort required by users to receive and act on system recommendations[44]. In order for a CDSS to fit the workflow of a particular clinic, customization of the CDSS might be necessary. Therefore, the customization functionality offered by each CDSS should be taken into account during selection[28]. Another consideration related to the local workflow is whether all the necessary data for the proper functioning of the CDSS is available in that specific point in the workflow.[45]

Another factor to consider when selecting a CDSS is its usability; more specifically how easy is it to use or how much training is needed to be able to use the CDSS. Vendors need to be clear about the expertise required for using the system.

An important consideration when selecting a CDSS should be its cost-effectiveness[46] compared to alternative CDSS or even other medical devices (e.g. a new piece of equipment). However, it remains difficult to demonstrate the return on investment of CDSS, especially against many competing priorities at the delivery system level. [34] A comprehensive assessment of the costs involved in the acquisition of a CDSS should be undertaken prior to its purchase, including one-off costs (purchase, training, implementation, etc.) but also costs incurred over time such as maintenance costs and resource utilization (e.g. time of its users). These costs should be weighed against not only estimated improvements in health outcomes but also estimated savings due to efficiencies facilitated by the CDSS.

Other factors to consider include the compatibility with legacy applications, the maturity of the CDSS, and the availability of upgrades. [23]

3.0 ACCEPTANCE TESTING

For the purpose of acceptance testing, a CDSS can best be seen a medical device for which many processes are already usually in place in health care providers. Acceptance tests for a medical device assure that the all defined specifications are fulfilled and that the medical device satisfies pertinent safety requirements.[47] These tests are usually defined by the CDSS vendor, but should be run in the presence of the representatives of the local site. On successful completion of the acceptance tests, the acceptance report will be signed and the payment for the device approved. Consequently, the set of test cases should be comprehensive, including covering cases on the edge of the domain of the CDSS, usually termed corner cases. The technical aspects of acceptance tests should be conducted by technology representatives while tests focused at usability or clinically oriented tests should be conducted by a subgroup of users that comprises a representative sample of the intended end-user population.

The acceptance test plan should cover at least the following aspects:

1. Installation and setup of the device.

2. Proper functioning of APIs offered by the CDSS (if any).

3. A complete walkthrough of the user interface, operating the CDSS as part of the existing workflow.

4. Clinical completeness, relevance, comprehensibility, consistency and repeatability of the output of the CDSS.

5. Auditing, security and privacy functions.

6. Typical error scenarios, such as unexpected, incorrect or incomplete input data, abrupt closure scenarios (e.g. due to power outage) leading to incomplete transactions, etc. The CDSS should not output inappropriate recommendations in the event of incomplete or inaccurate data. Moreover, the CDSS is expected to handle these situations by keeping internal consistency, providing appropriate error messages and, if necessary, proceeding to an orderly shutdown.

In addition to the above, acceptance testing of a CDSS should test the accuracy of the CDSS recommendations, as inaccurate recommendations might endanger the safety or well-being of patients. These tests should compare the outcome of the CDSS to the expected outcome on a fixed, small and restricted but representative sample of real cases. The estimated accuracy based on these acceptance test results should be compared against the accuracy claimed by the vendor and statistically test whether it is within the specified error tolerance. The same applies to the other quantitative and qualitative estimates provided by the vendor. In order to test whether the real accuracy of the CDSS (or any other parameter) is within a given error tolerance based on a sample of tests, a statistical test (e.g. Mann-Whitney U test) should be used to calculate the probability that the accuracy observed in the sample belongs to a probability distribution determined by the claimed accuracy and error tolerance. If the calculated probability is below a certain significance threshold, we can reject the hypothesis that the actual accuracy is within the error tolerance. Finally, a check for completeness and accessibility of the CDSS user manual as part of acceptance testing would be important for novice users or in emergency, unusual situations.

4.0 COMMISSIONING

Commissioning is the process that prepares the CDSS for safe clinical use in the local site, meeting established requirements and end users' expectations.[48] As such, commissioning verifies that the CDSS has been installed in the local site following the agreed requirements, successfully handed over from the vendor, and most importantly, that it functions properly. It is widely recommended to prepare for this phase by devising a commissioning plan that describes the tasks, schedule, and required human and equipment resources as well as the amount of support required from the CDSS vendor.

The first step in the commissioning plan is installation in the local site, which in the case of CDSSs, inevitably requires some degree of configuration or customization. Customization might be required for technical or safety reasons, for example to make sure that parameters in the CDSS are correctly linked to the local EHR and that definitions of clinical terms are in sync between the CDSS and local EHR. Customization is also a powerful tool to make the output of the CDSS more relevant, useful and safe for use. [39] A qualitative study found that all successful sites devoted considerable staff time to customization of their CDSS. [45] An example of customization could be to assess and improve the appropriateness of alerts to avoid alert fatigue. [10]

In order to test that the installed CDSS functions properly in the local site, a test plan needs to be designed and executed. To begin with, the implementation of the CDSS is likely to require some changes to the workflow on the users end. In that case, the information necessary to support the future workflow needs to be identified and the new workflow tested. Once the new workflow is established, the aim is to ensure the CDSS is functioning properly by testing as many clinically relevant scenarios and corner cases as possible. The steering committee formed by clinicians, administrators and technology experts should be involved in identifying all the relevant situations and corner cases where the installed CDSS could fail in the local site environment and lead to poor quality or reliability. A set of past cases, which includes difficult and rare cases along with a representative sample of the local case population could be retrospectively tested if a database with past cases exists. In this case, the recommendations of the CDSS is either assessed by a panel of clinical experts in a blind study where the experts ignore the CDSS's output or compared against the decisions that were taken in the

past. However, is important that CDSS should be tested on real world cases from the users' own clinical practice prior to implementation.[45] An option is to test the CDSS prospectively by running a pilot program where the CDSS is used in parallel to the existing workflow or where the CDSS is used with supervision using the existing workflow as fallback.[49] Strategies to cover a representative sample of usual and rare cases, include random sampling, input selection and control flow testing.[50] During the pilot, it is interesting to perform an initial assessment of the clinical relevance of the CDSS in terms of user acceptance, adherence to the CDSS's recommendations and its impact on the clinical decisions and ultimately on patient or health outcomes. Significant deviations on the estimated performance of the CDSS during this phases as compared with that in acceptance testing or vendor's claims of performance and error tolerance should be discussed with the vendor. Failure mode analysis is an important part of commissioning testing, where faults in data entry are simulated and the behavior or CDSS is analyzed and tested for consistency.[51] Testing during commissioning is also important to grow confidence of local physicians in that the support system works in their local setting. [13]

6.0 IMPLEMENTATION

The implementation process is another important factor in the success of a CDSS [52] and consists of the design and execution of the rollout plan, transitioning from the old workflow to the new one including the CDSS and the deployment of the CDSS within the local site. An effective implementation of CDSS requires preparing both users and the local site's infrastructure for the widespread use of the CDSS. The preparation of the infrastructure will vary across CDSSs and local sites but there are common themes on how to prepare users for the use of a new CDSS. Prior to and surrounding implementation, it is important to communicate with and educate the affected users[53]. Effective training of all the stakeholders and intended users of the CDSS is key to its success [54] and should comprise different aspects such as when (and when not) to use it, how to use it, how to interpret the output of the CDSS and when to override the CDSS recommendations, amongst others. It also includes helping users understand how the CDSS will impact their daily activities and how they can provide feedback.[53] It is important as part of the training to manage users' expectations in terms of efficiency and effectiveness and make sure users understand the strengths

and limitations of the CDSS. [22] Different stakeholders might have different expectations: some primarily view CDSSs as a vehicle for promoting standardization, quality, and safety while clinicians might see it differently.[45] Training should also serve the purpose of preparing users for a necessary leap of faith: a CDSS will only be used if it is perceived as beneficial by those using it, but the benefits of the CDSS will be appreciated only after overcoming the initial challenges of using it.[33] Hands-on training is a valuable tool, as users might need some handholding at first, as is on-site support from vendors as needed to help with any immediate issues that may occur. [53] The deployment or rollout of the CDSS can be undertaken incrementally, (e.g. rolling it out in a single post or facility to "get the kinks worked out") or all at once, which requires good preparation. [32]

7.0 QUALITY ASSURANCE

Before the CDSS has been deployed, it is crucial to design a quality assurance (QA) program to ensure that the performance and safety of the CDSS is maintained by assuring its quality remains fit for purpose throughout its life cycle.

As part of the QA program for a CDSS, performance must be defined using a set of metrics in terms of efficiency and efficacy so that the impact of the CDSS can be measured over time.[45] Measures of efficacy might be specific to the functioning of the CDSS (e.g. sensitivity and specify for a diagnostic tool) or generic, such as patient safety and change in health outcomes (such as life expectancy). Efficiency can be measured in resources saved, such as costs and productivity.[34] In order to assess the CDSS performance, it is especially valuable to quantify baseline performance levels (i.e. before the implementation of the CDSS), as well as have an estimate of the target performance upfront.[53]

The QA plan must guarantee that any malfunctions are identified and resolved in the shortest time possible. To facilitate the discovery of CDSS malfunctions, mechanisms need to be in place for receiving user feedback and acting on it.[55] Besides, CDSS malfunctions can be identified by a combination of qualitative and quantitative analyses (e.g. of firing rates for alert systems or overrides for recommender CDSS[28]. Visual detection and statistical process control analysis have shown good results as tools to detect malfunction.[56] In addition to malfunctions, it is important to log

or track the cases where the CDSS was not adhered to (such as when an alert was ignored or a recommendation overridden), as knowing how often the CDSS is being overridden and why can offer valuable insights and lead to identification of previously undetected malfunction. [45] Similarly, monitoring proper utilization of the installed CDSS is important as this could lead to a reduced performance.[22]

At the data quality front, local sites have to define and enforce internal standards to assure the integrity of entered data [45]. Data providers to the CDSS should be trained about the importance of high-quality data and their responsibility in assuring its accuracy.

The QA plan must also assure that the performance and safety of the CDSS is maintained over time. In this sense, a CDSS is not radically different from a treatment planning system or a radiotherapy linear accelerator, because deviations of CDSS performance beyond certain bounds of tolerance have the potential to cause medical mistreatment. For example, Nakatsugawa et al. [57] observed the need to update the prediction models with prospective data collection for maintaining the performance of their RT-induced toxicity prediction models. The first concern in this aspect is external (context) drift over time. Patterns of clinical practice are constantly evolving over time: changes in the clinic are sometimes radical (such as the introduction of image-guided radiation therapy or robotic surgery) and gradual at other times (e.g. percentage of patients with oropharyngeal squamous cell carcinoma expressing the p16 protein from human papilloma viral infection). Changes in patient case mixture, obsolescence of certain drugs and treatments, recoding of prognostic clinical features and clinical guidelines based on new randomized trials could all lead to unwanted divergence of CDSS recommendations over time. Such changes are often impossible to forecast during CDSS acceptance testing and commissioning and represent potential sources of time-dependent inconsistencies that violate the original assumptions built into the CDSS. These shifts can be related either to the input of the CDSS (e.g. clinical presentation of patients changing significantly since the CDSS was initially commissioned, thus exposing a previously unknown systematic bias towards certain patient subgroups) or its output (whereby the CDSS makes recommendations that are not in line with the most recent clinical guidelines). One other potential source of temporal divergence is internal (model) drift. The models underpinning the CDSS are likely not to remain static, but must be updated at specific times well after commissioning of the original CDSS. In addition, models developed

on limited sample sizes may initially incorporate some systematic bias that will be gradually reduced over time as the models are fed with progressively larger datasets on which to train and validate on. As described elsewhere,[58] models could be updated via any one of the following : (i) shifting either the baseline risk level or (for the case of binary models) the cut-off value for binary outcome, (ii) computing new values for an existing set of parameters or (iii) the model is trained afresh on expanded data, leading to possibly new model parameters, new coefficients and (for binary outcomes) new cut-off values.

A suitable safeguard for internal and external drift is to establish and routinely review incident monitoring logs for inappropriate or incorrect responses from the CDSS. At the same time, a 'repeated local validation' cohort should be assembled from time to time or preferably continuously to critically re-examine the tests done during the commissioning stage. The repetition may help to ensure that the CDSS remains clinically valid, despite changes in local practice or evidence based guidelines. Such a continuous local validation infrastructure will also be beneficial when introducing an update to the CDSS (see below). Finally, it is important to re-emphasise that no CDSS can ever be perfect, but at the very least the quality assurance system will document that the performance of the CDSS meets criteria based on the commission results as a benchmark.

Among the top priorities for the CDSS steering committee would be to establish an update management protocol. CDSS, in common with medical software in general, are most likely to be updated in the "offline" mode. That is, via a vendor-instigated or user-instigated change request, a CDSS is temporarily taken out of clinical use and placed in "maintenance" mode. Subsequent changes are performed in the maintenance state, such as applying a software version upgrade or correcting of faulty function. In analogy with other aspects of maintenance and QA of clinical systems, "clinical hand-over", i.e. acceptance of the system back into clinically operational mode, following any such update should only be allowed after some CDSS performance verification checks have been performed on the changed system. The minimum necessary tests should have been pre-specified by the vendor or the maintenance manual based on risk analysis, but it may be advisable to include a some additional tests taken from the acceptance testing procedures, in order to certify that all of the essential functionality of the CDSS has been restored following the update. With migration of medical software systems to "cloud services", increasing system automation and mathematical algorithms that

are able to learn "on-the-fly", one also has to countenance the possibility of CDSSs that update "online". Such CDSSs can be allowed to evolve in real time based on interactions between the user and its recommendations, such that the behavior of the CDSS might slightly change with each interaction. An update management protocol may explicitly permit online updates, which pose a new and interesting challenge, that of seeking the ideal trade-off between the potential of continuous improvement of performance against the risk of undetected performance degradation due to, for example, systematic biases in the input.

Another top priority should be to implement a routine QA test schedule that specifies which tests should be done, when they should be done and by whom.(53) As part of the QA tests, various aspects of the functionality of the CDSS are tested against an agreed upon ground truth. As a general rule, the types of QA tasks are drawn from the same checks as for commissioning. Therefore, the documented results of commissioning can be re-used at specified time intervals, in order to certify that the CDSS performance has not unduly drifted over time. Multiple statistical anomaly detection models applied to anomaly detection on CDSS over time have been described and compared in the literature, and the most appropriate method will depend on the nature of the CDSS.(59,60) The nature and frequency of such QA tests depend on the likelihood of unwanted deviation in CDSS performance and its potential consequences. QA tests should be performed more frequently for either highly likely failures or non-conformance events that lead to severe consequences. On the other hand, unlikely failures and events that do not have major clinical consequences need only to be checked infrequently. An important effort should be directed towards procedural mitigation of rare failures that carry severe consequences, because this may not be easy to intercept within a routine QA program.

5.0 CONCLUSION

CDSSs have shown great potential for improving healthcare and patient safety as well as reducing unwarranted variation, resource use and costs. However, an inaccurate or inappropriate CDSS might deteriorate the quality of healthcare and put patients at risk.[25] Therefore, considerable care must be taken to minimize the potential adverse consequences of CDSSs.[61] It is important to select carefully the CDSS that matches the requirements of the local site. As with any other medical device, CDSSs require stringent acceptance testing, commissioning and quality assurance by the local site[13]. In addition, an effective implementation plan is key to overcome barriers for a successful CDSS[62]. In the present review we have summarized the guidance collected from the literature in order to provide CDSS implementers. We conclude that following a systematic approach to the different aspects involved in the adoption of a CDSS will help avoid pitfalls, improve patient safety and increase the chances of success.

REFERENCES

1. *Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44.*
2. *Wyatt J, Spiegelhalter D. Evaluating Medical Expert Systems: What To Test, And How ? Talmon JL, Fox J, editors. Knowl Based Syst Med Methods Appl Eval. 1991;274–90.*
3. *Dombal FT de, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided Diagnosis of Acute Abdominal Pain. Br Med J. 1972 Apr 1;2(5804):9–13.*
4. *Shortliffe E. Computer-Based Medical Consultations: MYCIN - 1st Edition [Internet]. 1976 [cited 2018 Dec 4]. Available from: https://www.elsevier.com/books/computer-based-medical-consultations-mycin/shortliffe/978-0-444-00179-5*
5. *Kuperman GJ, Gardner RM, Pryor TA. HELP: A Dynamic Hospital Information System [Internet]. New York: Springer-Verlag; 1991 [cited 2019 Mar 1]. (Computers and Medicine). Available from: https://www.springer.com/gp/book/9781461277859*
6. *Wright A, Sittig DF. A four-phase model of the evolution of clinical decision support architectures. Int J Med Inf. 2008 Oct 1;77(10):641–9.*
7. *Pryor TA, Hripcsak G. The Arden syntax for medical logic modules. Int J Clin Monit*

Comput. 1993 Nov;10(4):215–24.

8. *Parker CG, Rocha RA, Campbell JR, Tu SW, Huff SM. Detailed clinical models for sharable, executable guidelines. Stud Health Technol Inform. 2004;107(Pt 1):145–8.*

9. *Wright A, Sittig DF, Ash JS, Feblowitz J, Meltzer S, McMullen C, et al. Development and evaluation of a comprehensive clinical decision support taxonomy: comparison of front-end tools in commercial and internally developed electronic health record systems. J Am Med Inform Assoc. 2011 May 1;18(3):232–42.*

10. *McCoy AB, Waitman LR, Lewis JB, Wright JA, Choma DP, Miller RA, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. J Am Med Inform Assoc. 2012 May 1;19(3):346–52.*

11. *Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S and Dean J. A guide to deep learning in healthcare. Nature Medicine. 2019;25:24–29.*

12. *Shortreed SM, Laber E, Lizotte DJ, Stroup TS, Pineau J, Murphy SA. Informing sequential clinical decision-making through reinforcement learning: an empirical study. Mach Learn. 2011 Jul 1;84(1–2):109–36.*

13. *Lambin P, van Stiphout RGPM, Starmans MHW, Rios-Velazquez E, Nalbantov G, Aerts HJWL, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol. 2013 Jan;10(1):27–40.*

14. *Valdes G, Simone CB, Chen J, Lin A, Yom SS, Pattison AJ, et al. Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2017;125(3):392–7.*

15. *Hemens BJ, Holbrook A, Tonkin M, Mackay JA, Weise-Kelly L, Navarro T, et al. Computerized clinical decision support systems for drug prescribing and management: A decision-maker-researcher partnership systematic review. Implement Sci. 2011 Aug 3;6(1):89.*

16. *Nieuwlaat R, Connolly SJ, Mackay JA, Weise-Kelly L, Navarro T, Wilczynski NL, et al. Computerized clinical decision support systems for therapeutic drug monitoring and dosing: A decision-maker-researcher partnership systematic review. Implement Sci. 2011 Aug 3;6(1):90.*

17. *Souza NM, Sebaldt RJ, Mackay JA, Prorok JC, Weise-Kelly L, Navarro T, et al. Computerized clinical decision support systems for primary preventive care: A decision-maker-researcher partnership systematic review of effects on process of care and patient outcomes. Implement Sci. 2011 Aug 3;6(1):87.*

18. *Roshanov PS, Misra S, Gerstein HC, Garg AX, Sebaldt RJ, Mackay JA, et al.*

Computerized clinical decision support systems for chronic disease management: A decision-maker-researcher partnership systematic review. Implement Sci. 2011 Aug 3;6(1):92.

19. *Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother Oncol. 2018 Feb 1;126(2):312–7.*
20. *Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes: A Systematic Review. JAMA. 1998 Oct 21;280(15):1339–46.*
21. *Cresswell K, Majeed A, Bates DW, Sheikh A. Computerised decision support systems for healthcare professionals: an interpretative review. J Innov Health Inform. 2013 Mar 22;20(2):115–28.*
22. *Berner ES, editor. Clinical Decision Support Systems: Theory and Practice [Internet]. 2nd ed. New York: Springer-Verlag; 2007 [cited 2018 Dec 3]. (Health Informatics). Available from: //www.springer.com/gp/book/9781441922236*
23. *Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review. JAMA. 2005 Mar 9;293(10):1223–38.*
24. *Bright TJ, Wong A, Dhurjati R, et al. Effect of clinical decision-support systems: A systematic review. Ann Intern Med. 2012 Jul 3;157(1):29–43.*
25. *Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc AMIA Symp. 2007 Oct 11;26–30.*
26. *Gooding MJ, Smith AJ, Tariq M, Aljabar P, Peressutti D, Stoep J van der, et al. Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test. Med Phys. 2018;45(11):5105–15.*
27. *Food and Drug Administration. Clinical and Patient Decision Support Software - Draft Guidance for Industry and Food and Drug Administration Staff [Internet]. 2018. Available from: https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm587819.pdf*
28. *Wright A, Ai A, Ash J, Wiesen JF, Hickman T-TT, Aaron S, et al. Clinical decision support alert malfunctions: analysis and empirically derived taxonomy. J Am Med Inform Assoc. 2018 May 1;25(5):496–506.*
29. *Eccles M, McColl E, Steen N, Nikki Rousseau, Grimshaw J, David Parkin, et al. Effect*

of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial. BMJ. 2002 Oct 26;325(7370):941.

30. *Keeffe B, Subramanian U, Tierney WM, Udris E, Willems J, Mcdonell M, et al. Provider Response to Computer-based Care Suggestions for Chronic Heart Failure. Med Care. 2005 May 1;43(5):461–5.*

31. *Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform. 2008 Apr 1;41(2):387–92.*

32. *Osheroff J, Teich J, Levick D, Saldana L, Velasco F, Sittig D, et al. Improving Outcomes with Clinical Decision Support: An Implementer's Guide, Second Edition [Internet]. CRC Press. [cited 2018 Dec 3]. Available from: https://www.crcpress.com/Improving-Outcomes-with-Clinical-Decision-Support-An-Implementers-Guide/Osheroff-Teich-Levick-Saldana-Velasco-Sittig-Rogers-Jenders/p/book/9780984457731*

33. *Coiera, E. Clinical Decision Support Systems. In: Guide to Health Informatics. 2003.*

34. *National Academy of Medicine. Optimizing Strategies for Clinical Decision Support [Internet]. [cited 2019 Jan 9]. Available from: https://nam.edu/optimizing-strategies-clinical-decision-support/*

35. *J Hummel. Integrating CDS Tools into Ambulatory Care Workflows for Improved Outcomes and Patient Safety | HealthIT.gov [Internet]. HealthIT.gov; [cited 2018 Dec 11]. Available from: https://www.healthit.gov/resource/integrating-cds-tools-ambulatory-care-workflows-improved-outcomes-and-patient-safety*

36. *Fox J, Thomson R. Clinical decision support systems: a discussion of quality, safety and legal liability issues. Proc AMIA Symp. 2002;265–9.*

37. *Redman TC. Data Driven: Profiting from Your Most Important Business Asset. Harvard Business Press; 2008. 273 p.*

38. *Musen MA, Middleton B, Greenes RA. Clinical Decision-Support Systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine [Internet]. London: Springer London; 2014 [cited 2018 Dec 3]. p. 643–74. Available from: https://doi.org/10.1007/978-1-4471-4474-8_22*

39. *Roshanov PS, Fernandes N, Wilczynski JM, Hemens BJ, You JJ, Handler SM, et al. Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. BMJ. 2013 Feb 14;346:f657.*

40. *Khorasani R, Hentel K, Darer J, Langlotz C, Ip IK, Manaker S, et al. Ten commandments for effective clinical decision support for imaging: enabling evidence-based practice to improve quality and reduce waste. AJR Am J Roentgenol. 2014 Nov;203(5):945–51.*

41. *Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018 Oct;2(10):719.*

42. *Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995 Dec;123(3):A12-13.*

43. *Campbell RJ. The Five Rights of Clinical Decision Support: CDS Tools Helpful for Meeting Meaningful Use. J AHIMA. 2013 Oct;84(10):42-47 (web version updated February 2016).*

44. *Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ. 2005 Mar 31;330(7494):765.*

45. *Ash JS, Sittig DF, Guappone KP, Dykstra RH, Richardson J, Wright A, et al. Recommended practices for computerized clinical decision support and knowledge management in community settings: a qualitative study. BMC Med Inform Decis Mak. 2012 Feb 14;12(1):6.*

46. *Jacob V, Thota AB, Chattopadhyay SK, Njie GJ, Proia KK, Hopkins DP, et al. Cost and economic benefit of clinical decision support systems for cardiovascular disease prevention: a community guide systematic review. J Am Med Inform Assoc. 2017 May 1;24(3):669–76.*

47. *Whelpton D. Acceptance testing of medical electrical equipment. J Med Eng Technol. 1984 Jan 1;8(1):19–23.*

48. *International Society for Pharmaceutical Engineering. Baseline Guide Volume 5: Commissioning & Qualification [Internet]. [cited 2019 Feb 12]. Available from: http://ispe.org/publications/guidance-documents/baseline-guide-volume-5-commissioning-qualification*

49. *Wagholikar KB, MacLaughlin KL, Kastner TM, Casey PM, Henry M, Greenes RA, et al. Formative evaluation of the accuracy of a clinical decision support system for cervical cancer screening. J Am Med Inform Assoc JAMIA. 2013 Jul;20(4):749–57.*

50. *Tso GJ, Yuen K, Martins S, Tu SW, Ashcraft M, Heidenreich P, et al. Test Case Selection in Pre-Deployment Testing of Complex Clinical Decision Support Systems. AMIA Summits Transl Sci Proc. 2016 Jul 20;2016:240–9.*

51. *Wexler A, Gu B, Goddu S, Mutic M, Yaddanapudi S, Olsen L, et al. FMEA of manual and automated methods for commissioning a radiotherapy treatment planning system. Med Phys. 2017;44(9):4415–25.*

52. *Connolly F, Byrne D, Lydon S, Walsh C, O'Connor P. Barriers and facilitators related to the implementation of a physiological track and trigger system: A systematic review of the qualitative evidence. Int J Qual Health Care. 2017 Dec 1;29(8):973–80.*

53. *Sirajuddin AM, Osheroff JA, Sittig DF, Chuo J, Velasco F, Collins DA. Implementation Pearls from a New Guidebook on Improving Medication Use and Outcomes with Clinical Decision Support. J Healthc Inf Manag. 2009;23(4):38–45.*

54. *Schuh C, de Bruin JS, Seeling W. Clinical decision support systems at the Vienna General Hospital using Arden Syntax: Design, implementation, and integration. Artif Intell Med. 2015 Dec 1;*

55. *Saleem JJ, Patterson ES, Militello L, Render ML, Orshansky G, Asch SM. Exploring Barriers and Facilitators to the Use of Computerized Clinical Reminders. J Am Med Inform Assoc. 2005 Jul 1;12(4):438–47.*

56. *Kassakian SZ, Yackel TR, Gorman PN, Dorr DA. Clinical decisions support malfunctions in a commercial electronic health record. Appl Clin Inform. 2017 Jul;8(3):910–23.*

57. *Nakatsugawa M, Cheng Z, Kiess A, Choflet A, Bowers M, Utsunomiya K, et al. The Needs and Benefits of Continuous Model Updates on the Accuracy of RT-Induced Toxicity Prediction Models Within a Learning Health System. Int J Radiat Oncol. 2019 Feb 1;103(2):460–7.*

58. *Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015 Jan 7;350:g7594.*

59. *Ray S, McEvoy DS, Aaron S, Hickman T-T, Wright A. Using statistical anomaly detection models to find clinical decision support malfunctions. J Am Med Inform Assoc. 2018 Jul 1;25(7):862–71.*

60. *Liu S, Wright A, Sittig DF, Hauskrecht M. Change-Point Detection for Monitoring Clinical Decision Support Systems with a Multi-Process Dynamic Linear Model. Proc IEEE Int Conf Bioinforma Biomed. 2017 Nov;2017:569–72.*

61. *Jenders RA. Advances in Clinical Decision Support: Highlights of Practice and the Literature 2015-2016. Yearb Med Inform. 2017 Aug;26(1):125–32.*

62. *Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci. 2017 Sep 15;12(1):113.*

CHAPTER 10

DISCUSSION AND FUTURE PROSPECTIVE

Geetha Mahadevaiah

In this chapter, a discussion of the key findings and lessons learned is provided. Also suggested are directions for future research and exploration.

Hospital workflows such as admission processes, emergency room operations, transfers to other hospitals, medication processes, medical documentation, supply of drugs, patient flow are still pain points in the smooth functioning of the hospitals. Workflow automation can improve the efficiency, time and more importantly patient satisfaction. The level of computerization in hospitals varies from one institution to another and from country to country. In the developing countries, the deployment of electronic medical record (EMR) systems and PACS systems is limited to a few private hospital chains. In some of the smaller hospitals, information technology is limited to patient registration and billing software. Thus, the valuable data needed for decision support is not available in digital form. Even in the top-notch hospitals of advanced countries, the digital data is incomplete or has missing semantic information and therefore not easily comprehended by computers for decision support.

The purpose of this thesis was to study these challenges and propose pragmatic solution by implementing the novel concepts as prototypes. As part of this thesis study, proof of concepts have been developed, to address the issues in the area of mining medical data, both imaging and text, fulfilling the stringent security and privacy requirements and the infrastructure needed for this technology. Further, Radiomics based proof points - classification of non-small lung cancer and fractals were developed. Lastly, the thesis discusses the issues and challenges in deploying a Clinical Decision Support System effectively in a hospital and provides recommendations.

A key finding is the need for an unobtrusive technology to provide the relevant information for the clinicians at the right time and place to make right decisions. Though medical informatics, diagnostics technologies are well established in the hospitals, the data generated is not yet fully exploited to improve hospital workflow, operations and patient outcomes. One of the reasons being that the data is not completely Findable, Accessible, Interoperable and Re-usable (FAIR) [11][1].The various characteristics to make the medical data FAIR compliant has been studied in-depth and successfully addressed as part of this thesis work in Chapters 2 to 5.

Though DICOM is a standard format for storing and accessing medical data, it is still vendor implementation dependent. The hierarchical nature of DICOM does not easily lend itself to data mining for decision support. In Chapters 2 to 5, the technologies

and concepts to make data findable and interoperable by leveraging the Semantic Web technologies and enabling consent and role based access to de-identified data for accessibility and re-use, were studied. The prototypes were developed using Radiotherapy DICOM data and lung cancer medical reports.

It was successfully demonstrated that the information from DICOM can be stored in the form of triples and natural language constructs can be used to query on the triples (RDF) data store (Chapter 5). Important medical information is stored in the verbose non-standard text formats of clinical reports. It is tedious for clinical practitioners to read and almost impossible for computers to find the context sensitive information from this unstructured data. The solution proposed in Chapter 2, leverages Natural Language Processing Techniques and ontologies to represent the information stored in the textual clinical report in a semantic graph. The graph representation improves readability for the clinical practitioners and provides the semantic meaning and computer comprehensibility, thus enabling automatic data mining.

The concerns and challenges of healthcare data security and privacy also impedes data sharing between hospital departments and across institutions. In Chapter 3, the latest technologies available for de-identification of Protected Health Information (PHI) were studied and the technology gap with respect to global requirements of PHI were identified. A solution to address this gap by increasing the robustness and accuracy of the NeuroNER models was developed. By this solution, the F-score of the MIMIC SpaCy model was improved from 67.5 to 97.4.

The security requirements of healthcare data stored in triples (RDF) format was explored and existing concepts discussed in the Chapter 4.A novel technique of building in consent with role-cum-rule based access to the triples data store was prototyped and successfully demonstrated.

In this thesis, the chapters 5 to 10, showcase how additional features or information can be extracted and derived from the widely available medical image data, to help clinicians in their decisions, includingthe infrastructure and commissioning strategies required for deploying an effective clinical decision support system in a hospital.

Radiomics offers immense possibilities to increase throughput and provide key information to support a clinician's decision making, such as tumor classification [15].A Radiomics based prototype successfully developed to automatically classify non small

lung cancer, is described in (Chapter 6). Similarly, Chapter 7 describes a methodology to derive a new feature called Fractal dimension to capture heterogeneity of the tumor based on tumor contour and shape. In this study, it was concluded that the Fractals based features play an important role in the histology classification of tumors.

In addition to the data, Radiomic case studies, the open technologies and infrastructure required to develop a system for medical image data analytics across multiple institutions was explored in Chapter 8.A successful prototype demonstration of storing de-identified PHI data in a cloud and applying Radiomics techniques for clinical decision support was achieved.

The primary requirement for data mining and Radiomics is to enable hospital throughput and support decision-making. The challenges of successfully deploying a CDSS in the hospital is elaborated and discussed in detailed in Chapter 9, concluding that good governance and deployment strategies for data and AI models in the hospitals, is a must for the effectiveness of the clinical decision support systems.

Future Prospective

In this section, a discussion on the challenges and future work in the areas of technologies to enable data analytics, FAIR data and unobtrusive computing is provided.

Technologies enabling Data Analytics

Radiomics is a medical field exploring deeper information from Radiology images to help in diagnosis and treatment. Cancer diagnosis and treatment often require a personalized analysis (bio-makers) for each patient, due to the heterogeneity among the different types of tumors and among patients. There are many treatment strategies and choosing the appropriate strategy is often based on personal bio-makers. Radiomics is a recent medical imaging scientific field that has lately shown encouraging results for achieving this personalization [5][6]. Semantically enriched meta-data serves as a basis for better data quality and provides theright options for feature extraction from medical images and textual data formats. Superior results are obtained from the higher precision of the prediction and classification calculated by machine learning algorithms. The proposed radiomics features looks at both

semantic data such as shape, size, volume, texture, intensity, location and syntactical data such as histograms, wavelets, fractals. This field has evolved from the oncology practice, where there is a need for combining data from various disciplines and reports to arrive at diagnoses and treatment strategies [14].

Cancer care also requires long term longitudinal clinical studies to track disease progression and effectiveness of treatment. In this regard, Radiomics plays an important role, as shown in the study of automatic classification of non-small lung cancer by extracting imaging features from annotated tumor regions. Similarly, in Chapter 7, the fractal dimension was used for the histology classification of tumors. The study showed thata fractals derived radiomics feature – max FD had a higher correlation with histology classification than other first order features. This study was conducted on 2D contour GTV regions only and may be extended to 3D using mesh approach increasing the accuracy of the Fractal Dimension. Also, the fractals study was limited to CT images, this may be extended to MR images and combined with other tissue characteristics derived from a MR image, to improve diagnosis and treatment decisions.

To facilitate comparison between and mining of information from various sources, the data extracted from radiology reports, pathology, genomics and histopathology can be stored in semantic web technologies format for ease of data mining [9] (Figure 1).

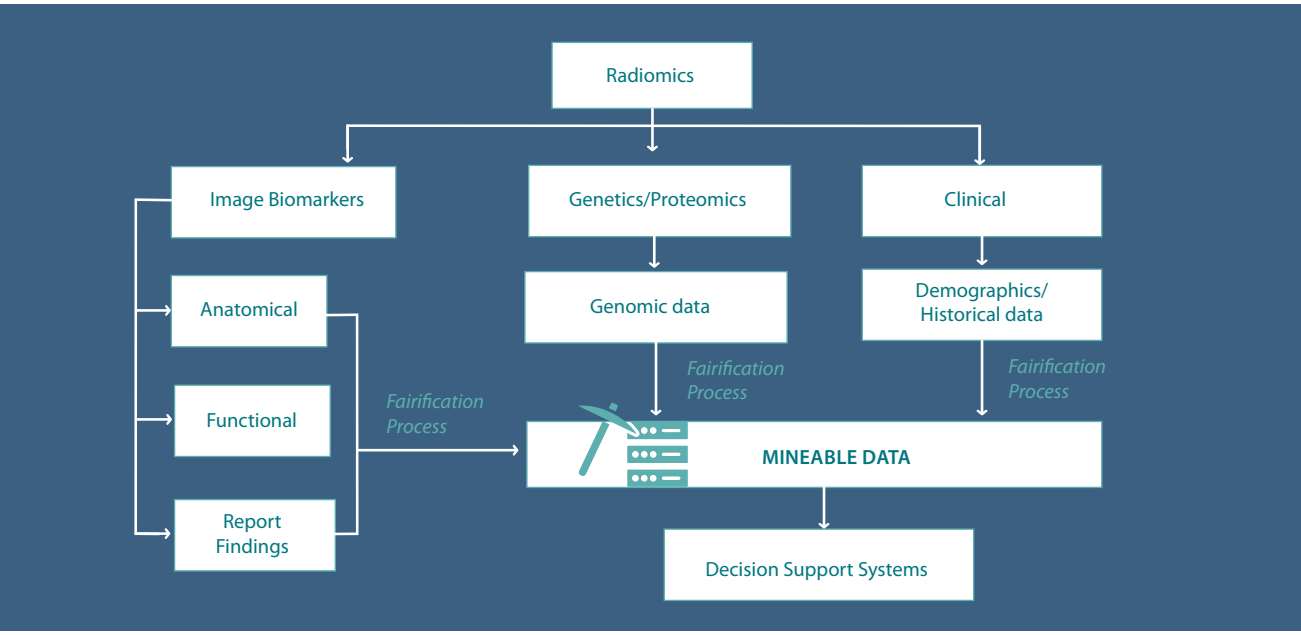


FIGURE 1. ***Radiomics Flow for medical data***

A semantic web data store is a collection of knowledge graphs in the form of RDF triples. Semantic web and its RDF data store provides a simple linked data model for data storage and sharing across institutions as described in Chapters2 and 5.

A major hurdle for data accessibility is the fragmented and diverse non-interoperable systems within a hospital and across institutions. There is a lack of standards and there are many proprietary systems, which make data sharing almost an impossible task. A reason for the lack of widespread adoption of the Semantic Web is the lack of security and control on the data stores, addressed in Chapter 4. Managing and tracking the changes in the knowledge graphs is a gap that can be addressed by storing these graphs in distributed ledgers like Blockchain [21], strengthening privacy and security.

Advances in Natural language processing (NLP) techniques can facilitate data mining in textual reports. In Chapter 2, NLP techniques were combined with ontologies and semantic web to automatically mine medical conditions from an unstructured medical report. In the proposed work, disease ontologies were developed for lung cancer. As a future work, these techniques can be extended to multiple organs and its corresponding cancer diseases. The current ontology structure, now defined for lung cancer, can also be extended to different cancer types and further to different disease types. Also, proposed techniques are designed based on traditional machine learning way of extracting features such as, Word2Vec, Term Frequency and the Inverse Document Frequency(TFIDF) and classifier training. To further improve the performance of the system, deep learning models like Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) can be tried.

Findable, Accessible, Interoperable and Re-usable Data

Radiomics, NLP, RDF are technologies, which lead to structured and processed data, usable by CDSS and learning systems [7]. AI and deep learning based decision support systems are rapidly gaining acceptance in the hospitals. The crux for the success of AI is availability of vast amounts of annotated data. Creation of successful deep learning algorithms is possible, if and only if large amounts of diverse, unprejudiced data with all the information is available.

Therefore, adherence to the FAIR principles for data management and curation is necessary. [23] The Figure 2 presents a visual representation of the FAIR guidelines for medical data. FAIR data would also help the AI community to validate and test the models on diverse data, thereby leading to higher accuracy of the results: "Good data stewardship is strategic to deep learning success".

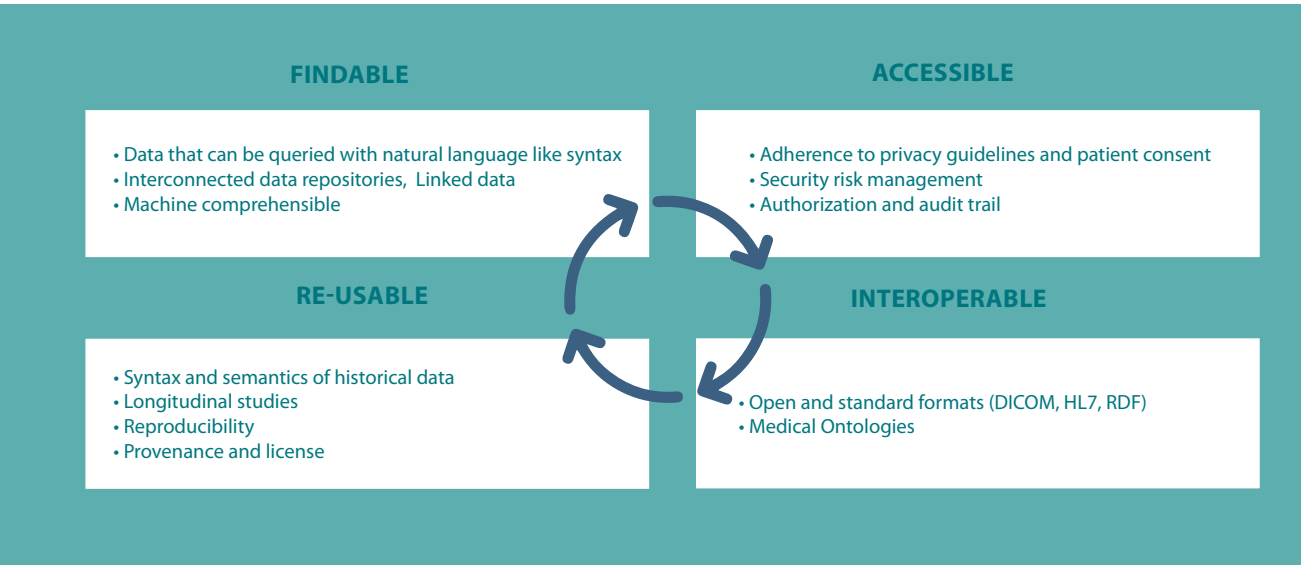


FIGURE 2. ***FAIR principles for medical data***

**Courtesy : Image Adapted from the Wikipedia : Implementing FAIR Data Principles - The Role of Libraries.pdf*

Protected health information (PHI) is identifiable personal information regulated by HIPAA or its equivalent in other jurisdictions. PHI examples are names, dates, address, social security number, license numbers, telephone numbers, emails, identifiable face from images etc.

De-identification tools and techniques are widely used to obfuscate the PHI information in medical data before it is made available to researchers and others [4]. Advances in medical and information technology has increased the PHI data availability and vulnerability. Hence, there is an almost constant evolution in techniques and tools to de-identify and protect the PHI information. The latest tools are based on artificial intelligence techniques, such as deep learning, to continuously learn PHI data and identify the same.

Interestingly during the study of the various techniques and tools for de-identification, it was noticed that the Indian names and other PHI information was not accurately recognized by some of the well-established techniques, such as NeuroNER, as discussed in Chapter 3. Given the diversity of ethnicity in the world, a much more robust algorithm is necessary for de-identification purposes.

Also, the process of de-identification of medical data is not rigorous and standardized across institutions. Due to this lacuna, there have been increasing cases of medical data inadvertently being made available without proper removal of PHI. This also results in data not being securely accessible for researchers and deep learning technologies. This is an area where medical care providers, researchers and vendors can bring in technology and standard practices to improve data accessibility.

Stricter privacy regulations and security aspects mandate that patient data does not leave the country, hence the cloud based solutions are restricted to a country or sometimes to just a hospital chain. In addition, hospitals also find it cumbersome, due to political and infrastructure related issues, to transfer even anonymized data outside their establishment.

Therefore, the concept of distributed learning wherein the local model learns on a local database, the local models then collaborate in a distributed platform to propose a global model based on negotiation and transfer learning approaches is gaining traction [17]. This global model is validated on the local databases of the participating sites [11][12]. This concept is being experimented by the 'Big Imaging data for Oncology in Netherlands and India Collaboration' (BIONIC) project [18]. The objective of the BIONIC project is to develop technology that allows global distributed learning and local deployment of clinical decision support systems(DSSs) based on Big Imaging Data analytics [2][13].

FAIR data is key enabler for a successful implementation of a distributed learning system. Widespread adoption of FAIR principles in storing medical data would increase the availability of data for learning systems [20].This concept of distributed learning approach and FAIR data stations will have a significant impact on the evolution and adoption of AI based models in hospitals.

Unobtrusive computing

The FAIR data and distributed learning approaches does not fully address a highly tenacious problem voiced by many clinicians, which is, the inherent requirement for clinicians to operate and feed data to computers.

The need of the hour, is efficient and effective data collection and curation methods and solutions. At the same time, clinicians and end users are bogged down by the large amount of data entry and reporting tasks. Startups and technology providers are experimenting with various tools to resolve the clinician's dichotomy. A few technology examples are electronic writing paper, speech recognition software supported by medical ontologies and intelligent bots to capture the information [16].

Another technology advancement that would alleviate some of this problem is the acceleration of hardware capabilities leading to availability of "AI on chips" in the near future. This would lead to availability of compute power everywhere, improving the workflow and interactions between clinicians and computers.

A fascinating advancement of AI technologies is to learn from less data using a top down approach, instead of the current bottom up, black-box approach of deep learning. In addition to transfer learning techniques, the new paradigm is to create AI models that are much more 'context aware'. For example: an AI model with prior knowledge of human anatomy can be more easily trained with less data, to make expert decisions in radiology.

The technology advances are rapidly changing the workflow and day-to-day operations in the hospital. The FDA guidelines and regulations are evolving to keep pace with the artificial intelligence based software providing clinical decision support [19]. As elaborated in Chapter 9, hospitals need to be better equipped with the right processes; training and quality measures to commission and deploy AI based decision support software.

In the past decade, great strides have been made towards addressing the myriad challenges in medical diagnostics and decision support systems. Today with the technological advancements, the delivery of healthcare is undergoing tremendous changes. Artificial intelligence, all pervasive computing, technologies such as block

chain are the game changers in all fields and more so in healthcare[22]. Healthcare industry , being a highly regulated industry, has been traditionally slow to accept newer technologies, but the need for cost effective value based care, improved patient outcomes, reducing stress and workload of clinicians has reached a tipping point and there is tremendous progress in the acceptance of the latest technology towards addressing these needs.

REFERENCES

1. *Go FAIR Initiative : https://www.go-fair.org/fair-principles/*
2. *Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT Timo M. Deist, A. Jochems ,Johan van Soest, Georgi Nalbantov, Cary Oberije, Seán Walsh ,Michael Eble , Paul Bulens , Philippe Coucke , Wim Dries, Andre Dekker, Philippe Lambin.*
3. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback*
4. *Hugues Duffau, Peggy Gatignol, Emmanuel Mandonnet et. al, "New insights into the anatomo-functional connectivity of the semantic system: a study using cortico-subcortical electrostimulations", Vol. 128, pg, 797-810, Brain, 2005*
5. *Lambin, Philippe, et al. "Radiomics: extracting more information from medical images using advanced feature analysis." European journal of cancer 48.4 (2012): 441-446.*
6. *Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJ, Dekker A, Fenstermacher D, Goldgof DB. Radiomics: the process and the challenges. Magnetic resonance imaging. 2012 Nov 1;30(9):1234-48.*
7. *Lambin P, Van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, Roelofs E, Van Elmpt W, Boutros PC, Granone P, Valentini V. Predicting outcomes in radiation oncology—multifactorial decision support systems. Nature reviews Clinical oncology. 2013 Jan;10(1):27.*
8. *Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, Carvalho S, Leijenaar RT, Nalbantov G, Oberije C, Marshall MS. Rapid Learning health care in oncology'–an approach towards decision support systems enabling customised radiotherapy. Radiotherapy and Oncology. 2013 Oct 1;109(1):159-64.*

9. *Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiotherapy and Oncology. 2013 Jul 1;108(1):174-9.*

10. *Skripcak T, Belka C, Bosch W, Brink C, Brunner T, Budach V, Büttner D, Debus J, Dekker A, Grau C, Gulliford S. Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets. Radiotherapy and Oncology. 2014 Dec 1;113(3):303-9.*

11. *Jochems, A., Deist, T.M., Van Soest, J., Eble, M., Bulens, P., Coucke, P., Dries, W., Lambin, P. and Dekker, A., 2016. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept. Radiotherapy and Oncology, 121(3), pp.459-467.*

12. *Dekker A, Vinod S, Holloway L, Oberije C, George A, Goozee G, Delaney GP, Lambin P, Thwaites D. Rapid learning in practice: A lung cancer survival decision support system in routine patient care data. Radiotherapy and Oncology. 2014 Oct 1;113(1):47-53.*

13. *Damiani A, Vallati M, Gatta R, Dinapoli N, Jochems A, Deist T, Van Soest J, Dekker A, Valentini V. Distributed learning to protect privacy in multi-centric clinical studies. In Conference on artificial intelligence in medicine in Europe 2015 Jun 20 (pp. 65-75). Springer, Cham.*

14. *Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. International Journal of Radiation Oncology* Biology* Physics. 2018 Nov 15;102(4):1143-58.*

15. *Jun Wang, Xia Liu, Di Dong, Jiangdian Song, Min Xu, Yali Zang, Jie Tian. Prediction of malignant and benign lung tumor using a quantitative radiomic method. 38th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC), 2016.*

16. *Tang PC, Patel VL. Major issues in user interface design for health professional workstations: summary and recommendations. International journal of bio-medical computing. 1994 Jan 1;34(1-4):139-48.*

17. *https://www.dtls.nl/fair-data/personal-health-train/*

18. *[18] https://www.nwo.nl/en/news-and-events/news/2015/ew/indo-dutch-collaboration-in-innovative-ict-research-further-strengthened.html*

19. *https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device*

20. *Kubben Pieter, Dumontier Michel, Dekker Andre, Fundamentals of Clinical Data Science, Springer Open, 2019.ISBN 978-3-319-99712-4 ISBN 978-3-319-99713-1 (eBook)*

21. *https://doi.org/10.1007/978-3-319-99713-1*

22. *Suveen Angraal,Harlan M. Krumholzand Wade L. Schulz Blockchain Technology Applicationsin Health Care, Cardiovascular Quality and Outcomes. 2017;10*

23. *https://doi.org/10.1161/CIRCOUTCOMES.117.003800*

24. *Advances in Computational Intelligence, Springer Science and Business Media LLC, 2019*

25. *Wilkinson MD, Verborgh R, Bonino da Silva Santos LO, Clark T, Swertz MA, Kelpin FDL, Gray AJG, Schultes EA, van Mulligen EM, Ciccarese P, Kuzniar A, Gavai A, Thompson M, Kaliyaperumal R, Bolleman JT, Dumontier M.2017. Interoperability and FAIRness through a novel combination of Web technologies. PeerJ Computer Science3:e110https://doi.org/10.7717/peerj-cs.110*

CHAPTER 11

SUMMARY

Geetha Mahadevaiah

Industrialization and automation of healthcare are crucial to transform the healthcare industry towards cost effectiveness, better outcomes, improved patient and staff satisfaction. With the high availability of cloud infrastructure, network bandwidth, cost efficient hardware and software technologies such as Semantic Web, AI, deep learning, this transformation is rapidly progressing. In this research, the role of Semantic Web technologies and machine learning in improving the clinical decision making for cancer care was studied. Chapter 2 elaborates on a method of using semantic technologies and machine learning to automatically detect cancer cases and classify cancer types, from unstructured medical reports.

The requirement of data privacy and security for healthcare data limits the sharing and accessing of data for research and clinical purposes. In Chapter 3, state of the art techniques for de-identification of patient notes in electronic health records were analyzed for their performance and a machine learning technique NeuroNER was studied for applicability of to Indian PHI data. This technique was further improved for reliability and performance, by the inclusion of a new model trained on Indian PHI data. Various methods addressing the security aspects of medical data stored on the Semantic Web was studied. This lead to the design of an innovative framework, described in Chapter 4, to manage the data access with the appropriate rights and required patient consent for the data stored in Semantic Web format (RDF).

In addition to the large amounts of textual data stored in medical records and reports, there is valuable and vast data stored in medical images, such as CT, MR and Radiotherapy. DICOM format is hierarchical and is not easy to mine. Chapter 5 describes a Semantic technology based architecture to extract clinically relevant information from the Radiotherapy DICOM data and make it easily findable using natural language like queries.

As a subsequent next step, Radiomics techniques were explored to increase the clinical findings from DICOM data. Radiomics is an evolving medical field for discovering deeper information from Radiology images to aid diagnosis and treatment. Chapter 6 explains a method for automating the classification of non-small lung tumors based on histopathology using Radiomics. Chapter 7 shows that Radiomics techniques can be extended further to gain deeper clinical insights by computing the Fractal dimension of a tumor to classify non small lung tumors. Fractals can be an important Radiomics feature, providing information not only about the Gross Tumor Volume (GTV) structure, but also help in characterization of the tumor.

Finally, system level needs and techniques to facilitate effective deployment and usage of clinical decision support in hospitals was studied. In Chapter 8, a system level architecture of a Cloud solution for image analytics, that can deployed in a hospital, is described. Chapter 9 provides an overview of the selection, acceptance, commissioning and quality assurance processes required in hospitals for effective deployment of machine learning based clinical decision support systems.

In conclusion, Chapter 10 provides a synopsis of the thesis and discusses the future advancements in the field of automation for healthcare.

SAMENVATTING

Industrialisering en automatisering zijn essentieel om de gezondheidszorg kosteneffectiever te maken, de resultaten te verbeteren en de tevredenheid van patiënten en personeel te vergroten. Dankzij de grote beschikbaarheid van technologische middelen, zoals de cloud-infrastructuur, netwerkbandbreedte, kostenefficiënte hardware en software zoals het semantische web, AI en deep learning, vindt deze transformatie in hoog tempo plaats. In dit onderzoek werd gekeken hoe semantische-webtechnologieën en automatisch leren kunnen worden benut om de klinische besluitvorming met betrekking tot de zorg bij kanker te verbeteren. In hoofdstuk 2 wordt uitleg gegeven over een methode waarbij semantische technologieën en automatisch leren worden gebruikt voor de automatisch detectie van gevallen van kanker en de classificering van vormen van kanker op basis van niet-gestructureerde medische dossiers.

Als gevolg van de wet- en regelgeving met betrekking tot gegevensbescherming en de beveiliging van zorggegevens kunnen deze gegevens slechts beperkt worden gedeeld en gebruikt voor onderzoek en klinische doeleinden. In hoofdstuk 3 wordt geanalyseerd hoe goed geavanceerde technieken voor de anonimisering van patiëntgegevens in elektronische dossiers werken en wordt gekeken naar de toepassing van de NeuroNER-techniek voor automatisch leren op privacygevoelige gegevens uit India. De betrouwbaarheid en prestaties van deze techniek werden verder verbeterd door een nieuw model toe te voegen dat werd ontwikkeld aan de hand van de gegevens uit India. Er werd gekeken naar verschillende methoden om om te gaan met de beveiligingsaspecten van medische gegevens die op het semantische web zijn opgeslagen. Op basis hiervan werd een innovatief raamwerk ontwikkeld waarmee de toegang tot gegevens opgeslagen in RDF (het formaat om gegevens te delen via het semantische web) kan worden beheerd, rekening houdend met de benodigde rechten en toestemming van de patiënt.

Er zijn niet alleen enorm veel gegevens in tekstvorm opgeslagen in medische dossiers en rapporten, maar ook in medische beelden, zoals CT-, MR- en röntgenbeelden is gigantisch veel waardevolle informatie vastgelegd. Maar de DICOM-indeling is hiërarchisch en kan niet eenvoudig doorzocht worden. In hoofdstuk 5 wordt een architectuur op basis van semantische technologie beschreven waarmee klinisch

relevante informatie kan worden geëxtraheerd uit radiotherapeutische DICOM-gegevens en eenvoudig gezocht kan worden met bijvoorbeeld zoekopdrachten.

Als volgende stap werd gekeken of Radiomics technieken konden worden toegepast om meer klinische informatie uit DICOM-gegevens te verkrijgen. Radiomics is een wetenschapsgebied in ontwikkeling dat erop gericht is om meer informatie te extraheren uit radiologische beelden ter ondersteuning van de diagnose en behandeling. In hoofdstuk 6 wordt een methode uitgelegd voor het automatiseren van de classificatie van niet-kleincellige longkanker (NSCLC) op basis van histopathologie met behulp van Radiomics technieken. In hoofdstuk 7 wordt aangetoond dat de toepassing van Radiomics technieken verder kan worden uitgebreid om diepergaande klinische inzichten te verwerven door de fractale dimensie van een tumor te berekenen voor de classificering van NSCLC. Fractals kunnen een belangrijke functie hebben bij Radiomics technieken; niet alleen geven ze informatie over het 'gross tumor volume' (GTV), maar daarnaast kunnen ze helpen bij het bepalen van de kenmerken van de tumor.

Tot slot werd er gekeken naar behoeften en technieken op ziekenhuisniveau om de effectieve toepassing van ondersteuning bij klinische besluitvorming in ziekenhuizen te vereenvoudigen. In hoofdstuk 8 wordt een architectuur op systeemniveau beschreven, gebaseerd op een cloudoplossing voor beeldanalyse die in ziekenhuizen kan worden toegepast. In hoofdstuk 9 wordt een overzicht gegeven van de processen voor selectie, aanvaarding, inbedrijfstelling en kwaliteitsborging die nodig zijn in ziekenhuizen voor een effectieve toepassing van automatisch lerende systemen ter ondersteuning van de klinische besluitvorming.

Afsluitend wordt er in hoofdstuk 10 een samenvatting gegeven van het proefschrift en worden de toekomstige ontwikkelingen op het gebied van automatisering in de gezondheidszorg besproken.

VALORIZATION ADDENDUM

The societal trends such as an ageing population, rising costs of healthcare expenditure, lack of skilled healthcare workers is forcing the healthcare industry to adopt newer solutions improving affordability and access to care. The availability of the right technologies is facilitating this transformation [1].

Royal Philips is a leading health technology company focused on improving people's health and enabling better outcomes across the health continuum from healthy living and prevention, to diagnosis, treatment and home care. The vision of Philips is to improve the lives of 3 billion people a year by 2030 [2].

As a Philips employee, the work on the thesis has influenced and contributed to the products and solutions as described in the paragraphs below.

Knowledge Dissemination

In addition to the knowledge sharing by publishing papers, the concepts and software developed as part of the thesis were shared amongst researchers at Maastricht University. The RDF graphs created for the work on Semantic representation of Radiotherapy Data for effective data mining and the RDF code created in Authorization Framework for Medical data was shared with the SEDI project team at Maastricht University, as a cross-sharing and learning initiative.

Chapter 7, Role of Fractals in histology classification for non-small lung cancer, is published as book chapter for the scientific community to follow. Classification based on fractals are expected to identify tumor habitats within the gross tumor volume. Finding of these habitats would be useful in targeted therapy for better prognosis.

Economical Exploitation

This thesis has contributed to solutions addressing the needs of patients, hospitals, and research institutes. A health technology company such as Philips Healthcare can valorize the following results of this thesis.

1. Data mining and semantic representation of clinical reports can be very useful for the clinical community for diagnosis and treatment pathways. The concepts and techniques proposed in chapters 2,4 and 5 can be part of a Philips platform providing the necessary infrastructure to mine clinical insights while adhering to strict privacy and security guidelines.

2. The de-identification models developed in Chapter 3, for text de-identification shall be part of the latest version of Philips HealthSuite Digital Platform (HSDP) [3].

3. The Radiomics models developed in chapter 6 and 7, automatic classification of tumor histopathology and Fractal Analysis for non-small lung cancer, are currently being verified and validated as part of the Philips Translation Research platform – IntelliSpace Discovery [4]. Based on the outcome of the validation phase, the models could become part of the Philips IntelliSpace Portal [5], in the near future.

4. Cloud based distributed learning for model training is gaining importance due to data privacy and security guidelines as in the latest EU General Data Protection Regulation GDPR [6]. The proposed research concepts in Chapter 8, "Cloud based Big data platform for image analytics" can be part of a Philips cloud-based solution for distributed learning.

Societal Expectation Management

In addition to likely economic benefit and knowledge sharing, the work done in this thesis has a larger influence on society. The technology to share the relevant clinical insights between hospitals across the globe, while adhering to privacy and security guidelines, shall enable researchers and clinicians to collaborate seamlessly and improve disease diagnosis and treatment at a rapid pace.

At the primary care level, the cloud-based platform for image analytics, described in Chapter 8, can provide telemedicine capabilities, connecting the experienced radiologists practicing in the large cities to physicians in remote villages and towns [7].

Similarly, the clinical decision support systems deployed on a cloud-based platform can empower physicians and healthcare workers in primary care to improve their diagnosis and treatment strategies.

REFERENCES

1. *2019 Global health care outlook-Shaping the future. https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/gx-lshc-hc-outlook-2019.pdf*
2. *https://www.philips.com/a-w/about/company.html*
3. *HSDP - https://www.philips.co.in/healthcare/innovation/about-health-suite*
4. *IntelliSpace Discovery : https://philipsproductcontent.blob.core.windows.net/assets/20190515/509c084a1c25484380c9aa4e00cb6773.pdf*
5. *IntelliSpace Portal : https://www.philips.co.in/healthcare/product/HC881062/intellispace-portal-80-all-your-advanced-analysis-needs-one-comprehensive-solution*
6. *EU General Data Protection Regulation - https://eugdpr.org/*
7. *Thomas Bodenheimer and Hoangmai H. Pham. Primary Care: Current Problems and Proposed Solutions.Health AffairsVol. 29, No. 5: Reinventing Primary Care.*

Curriculum Vitae

Geetha Mahadevaiah was born on July 24th 1963 in Bangalore, India and her maiden name is Geetha Govindaraj. After completion of high school in 1982, she studied Computer Science and Engineering at Bangalore Institute of Technology, Bangalore University and graduated in Dec 1985. She started to work in the software industry from 1996. She studied 6th month certificate courses in Parallel Processing and Microprocessor based Design at Indian Institute of Science in 1987 and 1988, as part of the continuing education programme. In 1992, she enrolled in the part-time MBA evening course at Central College, Bangalore University and completed the Masters in Business Administration - Finance and Marketing in 1994.She joined Philips India Ltd., in 2000 in the healthcare. Here she grew interest and understanding of the medical domain. She moved to the Research department in Philips in the year 2010 and was fascinated by research work. This motivated her to enroll for the Phd studentship in Maastricht University, as an external industry candidate. The work done as part of her thesis on machine learning and semantic web for cancer care has influenced and benefited similar projects at Philips Research. She completed her thesis in 2019 under the able guidance of Prof. dr. ir. Andre Dekker and Dr. Leonard Wee.

CHAPTER 12

LIST OF PUBLICATIONS

1. Geetha Mahadevaiah, Dinesh M.S., Amogh Hiremath, Vani Agarwal, Ponnurangam Kumaraguru, Andre Dekker : AUTOMATING DATA MINING OF MEDICAL REPORTS: *International Journal of Computer Science and Technology (IJSCT) Vol.01, No.2, March 2019 DOI:0.5121/IJCST.2019.041012 1*

2. Geetha Mahadevaiah, Dinesh M.S, Rithesh Sreenivasan, Sana Moin and Andre Dekker[2] : DE-IDENTIFICATION OF PROTECTED HEALTH INFORMATION PHI FROM FREE TEXT IN MEDICAL RECORDS. *International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 8, No 1/2, May 2019*

3. Geetha Madadevaiah[1] , RV Prasad[1], Amogh Hiremath[1], Michel Dumontier[2], Andre Dekker[3] : AUTHORIZATION FRAMEWORK FOR MEDICAL DATA. International Journal of Database Management Systems (IJDMS) Vol.11, No.2/3, June 2019

4. Geetha Mahadevaiah, Johan Van Soest, Dr. Andre Dekker, Dr. Narendranath Udupa, Dr. Shyam Vasudev Rao , Y. Kiran Kumar , R.V.Prasad. SEMANTIC REPRESENTATION OF RADIOTHERAPY DATA FOR EFFECTIVE DATA MINING. *International Journal of Biomedical Science & Bioinformatics. Volume 3 : Issue 2 [ISSN 2475 - 2290] 31 August 2016.*

5. Ravindra Patil, Geetha Mahadevaiah and Andre Dekker. AN APPROACH TOWARD AUTOMATIC CLASSIFICATION OF TUMOR HISTOPATHOLOGY OF NON−SMALL CELL LUNG CANCER BASED ON RADIOMIC FEATURES. Tomography. 2016;2(4):374–377. doi:10.18383/j.tom.2016.00244

6. Ravindra Patil, Geetha Mahadevaiah, Srinidhi Bhat, Dinesh M.S, Leonard Wee and Andre Dekker. FRACTAL ANALYSIS IN HISTOLOGY CLASSIFICATION OF NON-SMALL CELL LUNG CANCER. *Book Chapter 4: Medical Imaging: Artificial Intelligence, Image Recognition, and Machine Learning. Page 63 https://books.google.co.in/books?hl=en&lr=&id=vHatDwAAQBAJ&oi=fnd&pg=PA63&ots=JLWaUQhX7b&sig=f2m5I-HCXyNXkf6-sKrv70TPI8o#v=onepage&q&f=false*

7. Sunil Kumar Vuppala, Dinesh M.S., Sreramkumar Viswanathan, Ganesan Ramachandran, Nagaraju Bussa and Geetha Mahadevaiah. CLOUD BASED BIG DATA PLATFORM FOR IMAGE ANALYTICS. *IEEE International Conference on Cloud Computing in Emerging Markets (CCEM) - 2017*

8. Geetha Mahadevaiah, RV Prasad , Inigo Bermejo, David Jaffray , Andre Dekker and Leonard Wee. MACHINE AND DEEP LEARNING BASED CLINICAL DECISION SUPPORT IN MODERN MEDICAL PHYSICS: SELECTION, ACCEPTANCE, COMMISSIONING AND QUALITY ASSURANCE. *Journal : Medical Physics, 10.1002/mp.13562, May 2019*

9. Ravindra Patil, Geetha Mahadevaiah, Leonard Wee, Andre Dekker. P2.01-052 DOES RADIOMICS IMPROVES THE SURVIVAL PREDICTION IN NON-SMALL CELL LUNG CANCER?. *Journal of Thoracic oncology, November 2017, Volume 12, Issue 11, Supplement 2, Page S2089. DOI: https://doi.org/10.1016/j.jtho.2017.09.1154*

Acknowledgments

I thank the members of the assessment and defense committee for their time and effort in assessing the dissertation thesis. I realize that significant time and effort has been devoted towards a review of the thesis as the topics cover various diverse areas in software technologies and clinical domain.

I am indebted to Prof. Dr. Andre Dekker for his encouragement and guidance towards completion of the thesis. Regardless of his very busy schedule, Dr. Andre spent effort and time for a thorough review providing invaluable feedback, teaching me to articulate my thoughts with clarity. Dr. Andre also involved me in external collaboration projects, such as the BIONIC and TRAIN. These projects provided avenues for experimentation, discussions and insights which fueled the topics of the thesis. I shall always cherish the deep and thought provoking discussions with Dr. Andre which has influenced not only the thesis but also my work. I hope to stay connected and continue to collaborate on exciting projects.

I thank my co-guide Dr. Leonard Wee for his guidance, candid feedback and comprehensive review of my thesis. Dr. Wee's approach to scientific work is truly inspiring. He is systematic, encourages open and deep discussions, which lead to higher quality. I look forward to continuing our association in projects spanning countries and institutions.

I am sincerely grateful to Dr. Shyam Vasudev Rao who was instrumental in motivating and encouraging me to enroll for the PhD studentship at Maastricht University. He has constantly monitored progress and inspired me towards achieving my goals.

I am very grateful to Prof. Dr. Jos Smits and other professors from Maastricht University who pioneered the Maastricht University's external student PhD. program in India. Dr. Jos Smits and his colleagues regularly visited our office in Philips India, reviewed the work of the PhD students and resolved administrative hurdles, if any. I thank Dr. Jos Smits for his words of encouragement and motivation.

I thank all my co-authors from Philips Research and Maastricht University. I received excellent support by means of in-depth review and re-work of the papers from my co-authors, Johan van Soest, Michel Dumontier, Inigo Bermejo and others from Maastricht University. I am grateful to my colleagues, Ritesh Sreenivasan, Dr. Sunil Kumar Vuppala, Nagaraju Bussa, Dr. Narendranath Udupa, Sana Moin, RV Prasad, Kiran Kumar Y, Amogh Hiremath, Vani Agarwal and others who have contributed in developing various topics and presenting them as papers.

Special thanks to Ravindra Patil, who is also working on his PhD at Maastricht University. We have collaborated together, discussed ideas, identified challenges etc., making the PhD journey a fun experience.

I am tremendously grateful to Dr. Dinesh M.S. who has helped me significantly in my work on the thesis. He has provided deep insights and frequent informal but thorough review of my work. His constant check on my progress and motivation saying ''it is easy, you can do it'' , greatly helped. Thank you, Dinesh !

I have great admiration for Dr. Hans Hofstraat's visionary leadership. He has inspired me to learn and delve deeper into oncology. I am extremely grateful for his encouragement and constant support.

I thank all my colleagues in Philips Research who have directly and indirectly helped me with my thesis. I have discussed various topics with many of them and it has aided me in gaining profound insights and broadened my thinking.

This would not have been possible without the moral support and help from my family. I am extremely grateful to all my family members for cheering me on and helping me fulfill my aspirations. Thank you !